Supervised Learning with Similarity Functions

Purushottam Kar
Indian Institute of Technology
Kanpur, INDIA
purushot@cse.iitk.ac.in

Prateek Jain
Microsoft Research Lab
Bangalore, INDIA
prajain@microsoft.com

Abstract

We address the problem of general supervised learning when data can only be accessed through an (indefinite) similarity function between data points. Existing work on learning with indefinite kernels has concentrated solely on binary/multi-class classification problems. We propose a model that is generic enough to handle any supervised learning task and also subsumes the model previously proposed for classification. We give a "goodness" criterion for similarity functions w.r.t. a given supervised learning task and then adapt a well-known landmarking technique to provide efficient algorithms for supervised learning using "good" similarity functions. We demonstrate the effectiveness of our model on three important supervised learning problems: a) real-valued regression, b) ordinal regression and c) ranking where we show that our method guarantees bounded generalization error. Furthermore, for the case of real-valued regression, we give a natural goodness definition that, when used in conjunction with a recent result in sparse vector recovery, guarantees a sparse predictor with bounded generalization error. Finally, we report results of our learning algorithms on regression and ordinal regression tasks using non-PSD similarity functions and demonstrate the effectiveness of our algorithms, especially that of the sparse landmark selection algorithm that achieves significantly higher accuracies than the baseline methods while offering reduced computational costs.

1 Introduction

The goal of this paper is to develop an extended framework for supervised learning with similarity functions. Kernel learning algorithms [1] have become the mainstay of discriminative learning with an incredible amount of effort having been put in, both from the theoretician's as well as the practitioner's side. However, these algorithms typically require the similarity function to be a positive semi-definite (PSD) function, which can be a limiting factor for several applications. Reasons being: 1) the Mercer's condition is a formal statement that is hard to verify, 2) several natural notions of similarity that arise in practical scenarios are not PSD, and 3) it is not clear as to why an artificial constraint like PSD-ness should limit the usability of a kernel.

Several recent papers have demonstrated that indefinite similarity functions can indeed be successfully used for learning [2, 3, 4, 5]. However, most of the existing work focuses on classification tasks and provides specialized techniques for the same, albeit with little or no theoretical guarantees. A notable exception is the line of work by [6, 7, 8] that defines a goodness criterion for a similarity function and then provides an algorithm that can exploit this goodness criterion to obtain provably accurate classifiers. However, their definitions are yet again restricted to the problem of classification as they take a "margin" based view of the problem that requires positive points to be more similar to positive points than to negative points by at least a constant margin.

In this work, we instead take a "target-value" point of view and require that target values of similar points be similar. Using this view, we propose a generic goodness definition that also admits the

goodness definition of [6] for classification as a special case. Furthermore, our definition can be seen as imposing the existence of a smooth function over a generic space defined by similarity functions, rather than over a Hilbert space as required by typical goodness definitions of PSD kernels.

We then adapt the landmarking technique of [6] to provide an efficient algorithm that reduces learning tasks to corresponding learning problems over a linear space. The main technical challenge at this stage is to show that such reductions are able to provide good generalization error bounds for the learning tasks at hand. To this end, we consider three specific problems: a) regression, b) ordinal regression, and c) ranking. For each problem, we define appropriate surrogate loss functions, and show that our algorithm is able to, for each specific learning task, guarantee bounded generalization error with polynomial sample complexity. Moreover, by adapting a general framework given by [9], we show that these guarantees do not require the goodness definition to be overly restrictive by showing that our definitions admit all good PSD kernels as well.

For the problem of real-valued regression, we additionally provide a goodness definition that captures the intuition that usually, only a small number of landmarks are influential w.r.t. the learning task. However, to recover these landmarks, the uniform sampling technique would require sampling a large number of landmarks thus increasing the training/test time of the predictor. We address this issue by applying a sparse vector recovery algorithm given by [10] and show that the resulting sparse predictor still has bounded generalization error.

We also address an important issue faced by algorithms that use landmarking as a feature constructions step viz [6, 7, 8], namely that they typically assume separate landmark and training sets for ease of analysis. In practice however, one usually tries to overcome paucity of training data by reusing training data as landmark points as well. We use an argument outlined in [11] to theoretically justify such "double dipping" in our case. The details of the argument are given in Appendix B.

We perform several experiments on benchmark datasets that demonstrate significant performance gains for our methods over the baseline of kernel regression. Our sparse landmark selection technique provides significantly better predictors that are also more efficient at test time.

Related Work: Existing approaches to extend kernel learning algorithms to indefinite kernels can be classified into three broad categories: a) those that use indefinite kernels directly with existing kernel learning algorithms, resulting in non-convex formulations [2, 3]. b) those that convert a given indefinite kernel into a PSD one by either projecting onto the PSD-cone [4, 5] or performing other spectral operations [12]. The second approach is usually expensive due to the spectral operations involved apart from making the method inherently transductive. Moreover, any domain knowledge stored in the original kernel is lost due to these task oblivious operations and consequently, no generalization guarantees can be given. c) those that use notions of "task-kernel alignment" or equivalently, notions of "goodness" of a kernel, to give learning algorithms [6, 7, 8]. This approach enjoys several advantages over the other approaches listed above. These models are able to use the indefinite kernel directly with existing PSD kernel learning techniques; all the while retaining the ability to give generalization bounds that quantitatively parallel those of PSD kernel learning models. In this paper, we adopt the third approach for general supervised learning problem.

2 Problem formulation and Preliminaries

The goal in similarity-based supervised learning is to closely approximate a *target* predictor $y : \mathcal{X} \to \mathcal{Y}$ over some domain $\mathcal{X}$ using a *hypothesis* $\hat{f}(\,\cdot\,; K) : \mathcal{X} \to \mathcal{Y}$ that restricts its interaction with data points to computing similarity values given by K. Now, if the similarity function K is not discriminative enough for the given task then we cannot hope to construct a predictor out of it that enjoys good generalization properties. Hence, it is natural to define the "goodness" of a given similarity function with respect to the learning task at hand.

Definition 1 (Good similarity function: preliminary)**.** *Given a learning task $y : \mathcal{X} \to \mathcal{Y}$ over some distribution $\mathcal{D}$, a similarity function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be (ϵ_0, B)-good with respect to this task if there exists some bounded weighing function $w : \mathcal{X} \to [-B, B]$ such that for at least a $(1 - \epsilon_0)$ $\mathcal{D}$-fraction of the domain, we have* $y(\mathbf{x}) = \mathop{\mathbb{E}}_{\mathbf{x}' \sim \mathcal{D}} \left[\!\left[w(\mathbf{x}') y(\mathbf{x}') K(\mathbf{x}, \mathbf{x}') \right]\!\right]$.

The above definition is inspired by the definition of a "good" similarity function with respect to classification tasks given in [6]. However, their definition is tied to class labels and thus applies only

Algorithm 1 Supervised learning with Similarity functions

Input: A target predictor $y : \mathcal{X} \to \mathcal{Y}$ over a distribution $\mathcal{D}$, an (ϵ_0, B)-good similarity function K, labeled training points sampled from $\mathcal{D}$: $\mathcal{T} = \left\{(\mathbf{x}_1^t, y_1), \ldots, (\mathbf{x}_n^t, y_n)\right\}$, loss function $\ell_S : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}^+$.

Output: A predictor $\hat{f} : \mathcal{X} \to \mathbb{R}$ with bounded true loss over $\mathcal{D}$

1: Sample d unlabeled landmarks from $\mathcal{D}$: $\mathcal{L} = \left\{\mathbf{x}_1^l, \ldots, \mathbf{x}_d^l\right\}$
`// Else subsample d landmarks from T (see Appendix B for details)`
2: $\Psi_{\mathcal{L}} : \mathbf{x} \mapsto 1/\sqrt{d}\left(K(\mathbf{x}, \mathbf{x}_1^l), \ldots, K(\mathbf{x}, \mathbf{x}_d^l)\right) \in \mathbb{R}^d$
3: $\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq B}{\arg\min} \sum_i^n \ell_S\left(\left\langle \mathbf{w}, \Psi_{\mathcal{L}}(\mathbf{x}_i^t)\right\rangle, y_i\right)$
4: **return** $\hat{f} : \mathbf{x} \mapsto \langle \hat{\mathbf{w}}, \Psi_{\mathcal{L}}(\mathbf{x})\rangle$

to classification tasks. Similar to [6], the above definition calls a similarity function K "good" if the target value $y(\mathbf{x})$ of a given point $\mathbf{x}$ can be approximated in terms of (a weighted combination of) the target values of the K-"neighbors" of $\mathbf{x}$. Also, note that this definition automatically enforces a smoothness prior on the framework.

However the above definition is too rigid. Moreover, it defines goodness in terms of violations, a non-convex loss function. To remedy this, we propose an alternative definition that incorporates an arbitrary (but in practice always convex) loss function.

Definition 2 (Good similarity function: final)**.** *Given a learning task $y : \mathcal{X} \to \mathcal{Y}$ over some distribution $\mathcal{D}$, a similarity function K is said to be (ϵ_0, B)-good with respect to a loss function $\ell_S : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ if there exists some bounded weighing function $w : \mathcal{X} \to [-B, B]$ such that if we define a predictor as $f(\mathbf{x}) := \mathop{\mathbb{E}}_{\mathbf{x}' \sim \mathcal{D}}[\![w(\mathbf{x}')K(\mathbf{x}, \mathbf{x}')]\!]$, then we have $\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}}[\![\ell_S(f(\mathbf{x}), y(\mathbf{x}))]\!] \leq \epsilon_0$.*

Note that Definition 2 reduces to Definition 1 for $\ell_S(a, b) = \mathbb{1}_{\{a \neq b\}}$. Moreover, for the case of binary classification where $y \in \{-1, +1\}$, if we take $\ell_S(a, b) = \mathbb{1}_{\{ab \leq B\gamma\}}$, then we recover the (ϵ_0, γ)-goodness definition of a similarity function, given in Definition 3 of [6]. Also note that, assuming $\sup_{\mathbf{x} \in \mathcal{X}} \{|y(\mathbf{x})|\} < \infty$ we can w.l.o.g. merge $w(\mathbf{x}')y(\mathbf{x}')$ into a single term $w(\mathbf{x}')$.

Having given this definition we must make sure that "good" similarity functions allow the construction of effective predictors (Utility property). Moreover, we must make sure that the definition does not exclude commonly used PSD kernels (Admissibility property). Below, we formally define these two properties and in later sections, show that for each of the learning tasks considered, our goodness definition satisfies these two properties.

2.1 Utility

Definition 3 (Utility)**.** *A similarity function K is said to be ϵ_0-useful w.r.t. a loss function $\ell_{actual}(\cdot, \cdot)$ if the following holds: there exists a learning algorithm $\mathcal{A}$ that, for any $\epsilon_1, \delta > 0$, when given poly$(1/\epsilon_1, \log(1/\delta))$ "labeled" and "unlabeled" samples from the input distribution $\mathcal{D}$, with probability at least $1 - \delta$, generates a hypothesis $\hat{f}(\mathbf{x}; K)$ s.t. $\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}}\left[\!\!\left[\ell_{actual}\left(\hat{f}(\mathbf{x}), y(\mathbf{x})\right)\right]\!\!\right] \leq \epsilon_0 + \epsilon_1$. Note that $\hat{f}(\mathbf{x}; K)$ is restricted to access the data solely through K.*

Here, the ϵ_0 term captures the *misfit* or the bias of the similarity function with respect to the learning problem. Notice that the above utility definition allows for learning from unlabeled data points and thus puts our approach in the semi-supervised learning framework.

All our utility guarantees proceed by first using unlabeled samples as *landmarks* to construct a landmarked space. Next, using the goodness definition, we show the existence of a good linear predictor in the landmarked space. This guarantee is obtained in two steps as outlined in Algorithm 1: first of all we choose d unlabeled landmark points and construct a map $\Psi : \mathcal{X} \to \mathbb{R}^d$ (see Step 1 of Algorithm 1) and show that there exists a linear predictor over $\mathbb{R}^d$ that closely approximates the predictor f used in Definition 2 (see Lemma 15 in Appendix A). In the second step, we learn a predictor (over the landmarked space) using ERM over a fresh labeled training set (see Step 3 of Algorithm 1). We then use individual task-specific arguments and Rademacher average-based generalization bounds [13] thus proving the utility of the similarity function.

2.2 Admissibility

In order to show that our models are not too rigid, we would prove that they admit good PSD kernels. The notion of a good PSD kernel for us will be one that corresponds to a prevalent large margin technique for the given problem. In general, most notions correspond to the existence of a linear operator in the RKHS of the kernel that has small loss at large margin. More formally,

Definition 4 (Good PSD Kernel)**.** *Given a learning task $y : \mathcal{X} \to \mathcal{Y}$ over some distribution $\mathcal{D}$, a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with associated RKHS $\mathcal{H}_K$ and canonical feature map $\Phi_K : \mathcal{X} \to \mathcal{H}_K$ is said to be (ϵ_0, γ)-good with respect to a loss function $\ell_K : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ if there exists $\mathbf{W}^* \in \mathcal{H}_K$ such that $\|\mathbf{W}^*\| = 1$ and*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_K\left(\frac{\langle \mathbf{W}^*, \Phi_K(\mathbf{x})\rangle}{\gamma}, y(\mathbf{x})\right)\right]\!\!\right] < \epsilon_0$$

We will show, for all the learning tasks considered, that every (ϵ_0, γ)-good PSD kernel, when treated as simply a similarity function with no consideration of its RKHS, is also $(\epsilon + \epsilon_1, B)$-good for arbitrarily small ϵ_1 with $B = h(\gamma, \epsilon_1)$ for some function h. To prove these results we will adapt techniques introduced in [9] with certain modifications and task-dependent arguments.

3 Applications

We will now instantiate the general learning model described above to real-valued regression, ordinal regression and ranking by providing utility and admissibility guarantees. Due to lack of space, we relegate all proofs as well as the discussion on ranking to the supplementary material (Appendix F).

3.1 Real-valued Regression

Real-valued regression is a quintessential learning problem [1] that has received a lot of attention in the learning literature. In the following we shall present algorithms for performing real-valued regression using non-PSD similarity measures. We consider the problem with $\ell_{\text{actual}}(a, b) = |a - b|$ as the true loss function. For the surrogates ℓ_S and ℓ_K, we choose the ϵ-insensitive loss function [1] defined as follows:

$$\ell_\epsilon(a, b) = \ell_\epsilon(a - b) = \begin{cases} 0, & \text{if } |a - b| < \epsilon, \\ |a - b| - \epsilon, & \text{otherwise.} \end{cases}$$

The above loss function automatically gives us notions of good kernels and similarity functions by appealing to Definitions 4 and 2 respectively. It is easy to transfer error bounds in terms of absolute error to those in terms of mean squared error (MSE), a commonly used performance measure for real-valued regression. See Appendix D for further discussion on the choice of the loss function.

Using the landmarking strategy described in Section 2.1, we can reduce the problem of real regression to that of a linear regression problem in the landmarked space. More specifically, the ERM step in Algorithm 1 becomes the following: $\mathop{\arg\min}_{\mathbf{w}\in\mathbb{R}^d : \|\mathbf{w}\|_2 \le B} \sum_i^n \ell_\epsilon(\langle \mathbf{w}, \Psi_{\mathcal{L}}(\mathbf{x}_i)\rangle - y_i)$.

There exist solvers (for instance [14]) to efficiently solve the above problem on linear spaces. Using proof techniques sketched in Section 2.1 along with specific arguments for the ϵ-insensitive loss, we can prove generalization guarantees and hence utility guarantees for the similarity function.

Theorem 5. *Every similarity function that is (ϵ_0, B)-good for a regression problem with respect to the insensitive loss function $\ell_\epsilon(\cdot, \cdot)$ is $(\epsilon_0 + \epsilon)$-useful with respect to absolute loss as well as $(B\epsilon_0 + B\epsilon)$-useful with respect to mean squared error. Moreover, both the dimensionality of the landmarked space as well as the labeled sample complexity can be bounded by $\mathcal{O}\left(\frac{B^2}{\epsilon_1^2}\log\frac{1}{\delta}\right)$.*

We are also able to prove the following (tight) admissibility result:

Theorem 6. *Every PSD kernel that is (ϵ_0, γ)-good for a regression problem is, for any $\epsilon_1 > 0$, $\left(\epsilon_0 + \epsilon_1, \mathcal{O}\left(\frac{1}{\epsilon_1\gamma^2}\right)\right)$-good as a similarity function as well. Moreover, for any $\epsilon_1 < 1/2$ and any $\gamma < 1$, there exists a regression instance and a corresponding kernel that is $(0, \gamma)$-good for the regression problem but only (ϵ_1, B)-good as a similarity function for $B = \Omega\left(\frac{1}{\epsilon_1\gamma^2}\right)$.*

3.2 Sparse regression models

An artifact of a random choice of landmarks is that very few of them might turn out to be "informative" with respect to the prediction problem at hand. For instance, in a network, there might exist *hubs* or *authoritative* nodes that yield rich information about the learning problem. If the relative abundance of such nodes is low then random selection would compel us to choose a large number of landmarks before enough "informative" ones have been collected.

However this greatly increases training and testing times due to the increased costs of constructing the landmarked space. Thus, the ability to prune away irrelevant landmarks would speed up training and test routines. We note that this issue has been addressed before in literature [8, 12] by way of landmark selection heuristics. In contrast, we guarantee that our predictor will select a small number of landmarks while incurring bounded generalization error. However this requires a careful restructuring of the learning model to incorporate the "informativeness" of landmarks.

Definition 7. *A similarity function K is said to be (ϵ_0, B, τ)-good for a real-valued regression problem $y : \mathcal{X} \to \mathbb{R}$ if for some bounded weight function $w : \mathcal{X} \to [-B, B]$ and* choice function *$R : \mathcal{X} \to \{0, 1\}$ with $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![R(\mathbf{x})]\!] = \tau$, the predictor $f : \mathbf{x} \mapsto \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[\![w(\mathbf{x}')K(\mathbf{x}, \mathbf{x}')|R(\mathbf{x}')]\!]$ has bounded ϵ-insensitive loss i.e. $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![\ell_\epsilon(f(\mathbf{x}), y(\mathbf{x}))]\!] < \epsilon_0$.*

The role of the choice function is to single out informative landmarks, while τ specifies the relative density of informative landmarks. Note that the above definition is similar in spirit to the goodness definition presented in [15]. While the motivation behind [15] was to give an improved admissibility result for binary classification, we squarely focus on the utility guarantees; with the aim of accelerating our learning algorithms via landmark pruning.

We prove the utility guarantee in three steps as outlined in Appendix D. First, we use the usual landmarking step to project the problem onto a linear space. This step guarantees the following:

Theorem 8. *Given a similarity function that is (ϵ_0, B, τ)-good for a regression problem, there exists a randomized map $\Psi : \mathcal{X} \to \mathbb{R}^d$ for $d = \mathcal{O}\left(\frac{B^2}{\tau\epsilon_1^2}\log\frac{1}{\delta}\right)$ such that with probability at least $1 - \delta$, there exists a linear operator $\tilde{f} : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle$ over $\mathbb{R}^d$ such that $\|\mathbf{w}\|_1 \leq B$ with ϵ-insensitive loss bounded by $\epsilon_0 + \epsilon_1$. Moreover, with the same confidence we have $\|\mathbf{w}\|_0 \leq \frac{3d\tau}{2}$.*

Our proof follows that of [15], however we additionally prove sparsity of $\mathbf{w}$ as well. The number of landmarks required here is a $\Omega(1/\tau)$ fraction greater than that required by Theorem 5. This formally captures the intuition presented earlier of a small fraction of dimensions (read landmarks) being actually relevant to the learning problem. So, in the second step, we use the *Forward Greedy Selection* algorithm given in [10] to learn a sparse predictor. The use of this learning algorithm necessitates the use of a different generalization bound in the final step to complete the utility guarantee given below. We refer the reader to Appendix D for the details of the algorithm and its utility analysis.

Theorem 9. *Every similarity function that is (ϵ_0, B, τ)-good for a regression problem with respect to the insensitive loss function $\ell_\epsilon(\cdot, \cdot)$ is $(\epsilon_0 + \epsilon)$-useful with respect to absolute loss as well; with the dimensionality of the landmarked space being bounded by $\mathcal{O}\left(\frac{B^2}{\tau\epsilon_1^2}\log\frac{1}{\delta}\right)$ and the labeled sampled complexity being bounded by $\mathcal{O}\left(\frac{B^2}{\epsilon_1^2}\log\frac{B}{\epsilon_1\delta}\right)$. Moreover, this utility can be achieved by an $\mathcal{O}(\tau)$-sparse predictor on the landmarked space.*

We note that the improvements obtained here by using the sparse learning methods of [10] provide $\Omega(\tau)$ increase in sparsity. We now prove admissibility results for this sparse learning model. We do this by showing that the dense model analyzed in Theorem 5 and that given in Definition 7 are interpretable in each other for an appropriate selection of parameters. The guarantees in Theorem 6 can then be invoked to conclude the admissibility proof.

Theorem 10. *Every (ϵ_0, B)-good similarity function K is also $\left(\epsilon_0, B, \frac{\bar{w}}{B}\right)$-good where $\bar{w} = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![|w(\mathbf{x})|]\!]$. Moreover, every (ϵ_0, B, τ)-good similarity function K is also $(\epsilon_0, B/\tau)$-good.*

Using Theorem 6, we immediately have the following corollary:

Corollary 11. *Every PSD kernel that is (ϵ_0, γ)-good for a regression problem is, for any $\epsilon_1 > 0$, $\left(\epsilon_0 + \epsilon_1, \mathcal{O}\left(\frac{1}{\epsilon_1\gamma^2}\right), 1\right)$-good as a similarity function as well.*

3.3 Ordinal Regression

The problem of ordinal regression requires an accurate prediction of (discrete) labels coming from a finite ordered set $[r] = \{1, 2, \ldots, r\}$. The problem is similar to both classification and regression, but has some distinct features due to which it has received independent attention [16, 17] in domains such as product ratings etc. The most popular performance measure for this problem is the absolute loss which is the absolute difference between the predicted and the true labels.

A natural and rather tempting way to solve this problem is to relax the problem to real-valued regression and threshold the output of the learned real-valued predictor using predefined thresholds $b_1, \ldots, b_r$ to get discrete labels. Although this approach has been prevalent in literature [17], as the discussion in the supplementary material shows, this leads to poor generalization guarantees in our model. More specifically, a goodness definition constructed around such a direct reduction is only able to ensure $(\epsilon_0 + 1)$-utility i.e. the absolute error rate is always greater than 1.

One of the reasons for this is the presence of the thresholding operation that makes it impossible to distinguish between instances that would not be affected by small perturbations to the underlying real-valued predictor and those that would. To remedy this, we enforce a (soft) margin with respect to thresholding that makes the formulation more robust to noise. More formally, we expect that if a point belongs to the label i, then in addition to being sandwiched between the thresholds b_i and b_{i+1}, it should be separated from these by a margin as well i.e. $b_i + \gamma \leq f(\mathbf{x}) \leq b_{i+1} - \gamma$.

This is a direct generalization of the margin principle in classification where we expect $\mathbf{w}^\top \mathbf{x} > b + \gamma$ for positively labeled points and $\mathbf{w}^\top \mathbf{x} < b - \gamma$ for negatively labeled points. Of course, wherein classification requires a single threshold, we require several, depending upon the number of labels. For any $x \in \mathbb{R}$, let $[x]_+ = \max\{x, 0\}$. Thus, if we define the γ-margin loss function to be $[x]_\gamma := [\gamma - x]_+$ (note that this is simply the well known hinge loss function scaled by a factor of γ), we can define our goodness criterion as follows:

Definition 12. *A similarity function K is said to be (ϵ_0, B)-good for an ordinal regression problem $y : \mathcal{X} \to [r]$ if for some bounded weight function $w : \mathcal{X} \to [-B, B]$ and some (unknown but fixed) set of thresholds $\{b_i\}_{i=1}^r$ with $b_1 = -\infty$, the predictor $f : \mathbf{x} \mapsto \mathop{\mathbb{E}}_{\mathbf{x}' \sim \mathcal{D}} [\![w(\mathbf{x}') K(\mathbf{x}, \mathbf{x}')]\!]$ satisfies $\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\left[f(\mathbf{x}) - b_{y(\mathbf{x})} \right]_\gamma + \left[b_{y(\mathbf{x})+1} - f(\mathbf{x}) \right]_\gamma \right] < \epsilon_0$.*

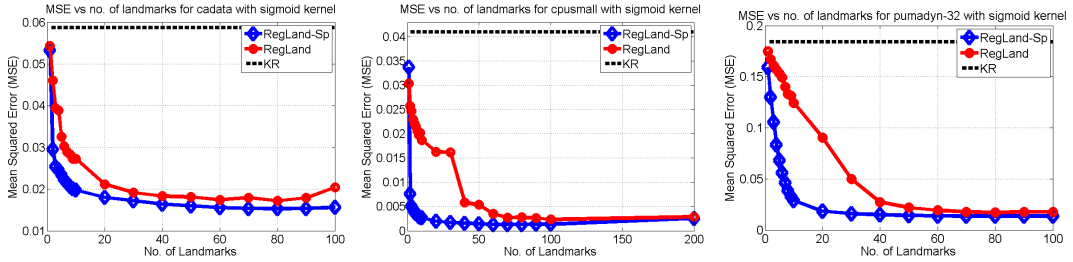
We now give utility guarantees for our learning model. We shall give guarantees on both the misclassification error as well as the absolute error of our learned predictor. We say that a set of points $x_1, \ldots, x_i \ldots$ is Δ-spaced if $\min_{i \neq j} \{|x_i - x_j|\} \geq \Delta$. Define the function $\psi_\Delta(x) = \frac{x + \Delta - 1}{\Delta}$.

Theorem 13. *Let K be a similarity function that is (ϵ_0, B)-good for an ordinal regression problem with respect to Δ-spaced thresholds and γ-margin loss. Let $\bar{\gamma} = \max\{\gamma, 1\}$. Then K is $\psi_{(\Delta/\bar{\gamma})}\left(\frac{\epsilon_0}{\bar{\gamma}}\right)$-useful with respect to ordinal regression error (absolute loss). Moreover, K is $\left(\frac{\epsilon_0}{\bar{\gamma}}\right)$-useful with respect to the zero-one mislabeling error as well.*

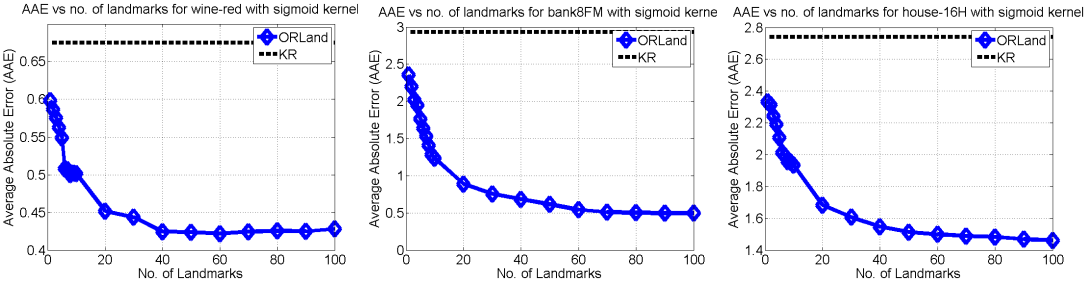
We can bound, both dimensionality of the landmarked space as well as labeled sampled complexity, by $\mathcal{O}\left(\frac{B^2}{\epsilon_1^2} \log \frac{1}{\delta}\right)$. Notice that for $\epsilon_0 < 1$ and large enough d, n, we can ensure that the ordinal regression error rate is also bounded above by 1 since $\sup_{x \in [0,1], \Delta > 0} (\psi_\Delta(x)) = 1$. This is in contrast with the direct reduction to real valued regression which has ordinal regression error rate bounded *below* by 1. This indicates the advantage of the present model over a naive reduction to regression.

We can show that our definition of a good similarity function admits all good PSD kernels as well. The kernel goodness criterion we adopt corresponds to the large margin framework proposed by [16]. We refer the reader to Appendix E.3 for the definition and give the admissibility result below.

Theorem 14. *Every PSD kernel that is (ϵ_0, γ)-good for an ordinal regression problem is also $\left(\gamma_1 \epsilon_0 + \epsilon_1, \mathcal{O}\left(\frac{\gamma_1^2}{\epsilon_1 \gamma^2}\right)\right)$-good as a similarity function with respect to the γ_1-margin loss for any $\gamma_1, \epsilon_1 > 0$. Moreover, for any $\epsilon_1 < \gamma_1/2$, there exists an ordinal regression instance and a corresponding kernel that is $(0, \gamma)$-good for the ordinal regression problem but only (ϵ_1, B)-good as a similarity function with respect to the γ_1-margin loss function for $B = \Omega\left(\frac{\gamma_1^2}{\epsilon_1 \gamma^2}\right)$.*



(a) Mean squared error for landmarking (RegLand), sparse landmarking (RegLand-Sp) and kernel regression (KR)



(b) Avg. absolute error for landmarking (ORLand) and kernel regression (KR) on ordinal regression datasets

Figure 1: Performance of landmarking algorithms with increasing number of landmarks on real-valued regression (Figure 1a) and ordinal regression (Figure 1b) datasets.

Datasets	Sigmoid kernel		Manhattan kernel	
	KR	Land-Sp	KR	Land-Sp
`Abalone` [18] $N = 4177$ $d = 8$	2.1e-002 (8.3e-004)	6.2e-003 (8.4e-004)	1.7e-002 (7.1e-004)	6.0e-003 (3.7e-004)
`Bodyfat` [19] $N = 252$ $d = 14$	4.6e-004 (6.5e-005)	9.5e-005 (1.3e-004)	3.9e-004 (2.2e-005)	3.5e-005 (1.3e-005)
`CAHousing` [19] $N = 20640$ $d = 8$	5.9e-002 (2.3e-004)	1.6e-002 (6.2e-004)	5.8e-002 (1.9e-004)	1.5e-002 (1.4e-004)
`CPUData` [20] $N = 8192$ $d = 12$	4.1e-002 (1.6e-003)	1.4e-003 (1.7e-004)	4.3e-002 (1.6e-003)	1.2e-003 (3.2e-005)
`PumaDyn-8` [20] $N = 8192$ $d = 8$	2.3e-001 (4.6e-003)	1.4e-002 (4.5e-004)	2.3e-001 (4.5e-003)	1.4e-002 (4.8e-004)
`PumaDyn-32` [20] $N = 8192$ $d = 32$	1.8e-001 (3.6e-003)	1.4e-002 (3.7e-004)	1.8e-001 (3.6e-003)	1.4e-002 (3.1e-004)

(a) Mean squared error for real regression

Datasets	Sigmoid kernel		Manhattan kernel	
	KR	ORLand	KR	ORLand
`Wine-Red` [18] $N = 1599$ $d = 11$	6.8e-001 (2.8e-002)	4.2e-001 (3.8e-002)	6.7e-001 (3.0e-002)	4.5e-001 (3.2e-002)
`Wine-White` [18] $N = 4898$ $d = 11$	6.2e-001 (2.0e-002)	8.9e-001 (8.5e-001)	6.2e-001 (2.0e-002)	4.9e-001 (1.5e-002)
`Bank-8` [20] $N = 8192$ $d = 8$	2.9e+000 (6.2e-002)	6.1e-001 (4.4e-002)	2.7e+000 (6.6e-002)	6.3e-001 (1.7e-002)
`Bank-32` [20] $N = 8192$ $d = 32$	2.7e+000 (1.2e-001)	1.6e+000 (2.3e-002)	2.6e+000 (8.1e-002)	1.6e+000 (9.4e-002)
`House-8` [20] $N = 22784$ $d = 8$	2.8e+000 (9.3e-003)	1.5e+000 (2.0e-002)	2.7e+000 (1.0e-002)	1.4e+000 (1.2e-002)
`House-16` [20] $N = 22784$ $d = 16$	2.7e+000 (2.0e-002)	1.5e+000 (1.0e-002)	2.8e+000 (2.0e-002)	1.4e+000 (2.3e-002)

(b) Mean absolute error for ordinal regression

Table 1: Performance of landmarking-based algorithms (with 50 landmarks) vs. baseline kernel regression (KR). Values in parentheses indicate standard deviation values. Values in the first columns indicate dataset source (in parentheses), size (N) and dimensionality (d).

Due to lack of space we refer the reader to Appendix F for a discussion on ranking models that includes utility and admissibility guarantees with respect to the popular NDCG loss.

4 Experimental Results

In this section we present an empirical evaluation of our learning models for the problems of real-valued regression and ordinal regression on benchmark datasets taken from a variety of sources [18, 19, 20]. In all cases, we compare our algorithms against kernel regression (KR), a well known technique [21] for non-linear regression, whose predictor is of the form:

$$f : \mathbf{x} \mapsto \frac{\sum_{\mathbf{x}_i \in \mathcal{T}} y(\mathbf{x}_i) K(\mathbf{x}, \mathbf{x}_i)}{\sum_{\mathbf{x}_i \in \mathcal{T}} K(\mathbf{x}, \mathbf{x}_i)}.$$

where $\mathcal{T}$ is the training set. We selected KR as the baseline as it is a popular regression method that does not require similarity functions to be PSD. For ordinal regression problems, we rounded off the result of the KR predictor to get a discrete label. We implemented all our algorithms as well as the

baseline KR method in Matlab. In all our experiments we report results across 5 random splits on the (indefinite) Sigmoid: $K(\mathbf{x}, \mathbf{y}) = \tanh(a \langle \mathbf{x}, \mathbf{y} \rangle + r)$ and Manhattan: $K(\mathbf{x}, \mathbf{y}) = -\|\mathbf{x} - \mathbf{y}\|_1$ kernels. Following standard practice, we fixed $r = -1$ and $a = 1/d_{\text{orig}}$ for the Sigmoid kernel where d_{orig} is the dimensionality of the dataset.

Real valued regression: For this experiment, we compare our methods (RegLand and RegLand-Sp) with the KR method. For RegLand, we constructed the landmarked space as specified in Algorithm 1 and learned a linear predictor using the LIBLINEAR package [14] that minimizes ϵ-insensitive loss. In the second algorithm (RegLand-Sp), we used the sparse learning algorithm of [10] on the landmarked space to learn the best predictor for a given sparsity level. Due to its simplicity and good convergence properties, we implemented the *Fully Corrective* version of the Forward Greedy Selection algorithm with squared loss as the surrogate.

We evaluated all methods using Mean Squared Error (MSE) on the test set. Figure 1a shows the MSE incurred by our methods along with reference values of accuracies obtained by KR as landmark sizes increase. The plots clearly show that our methods incur significantly lesser error than KR. Moreover, RegLand-Sp learns more accurate predictors using the same number of landmarks. For instance, when learning using the Sigmoid kernel on the `CPUData` dataset, at 20 landmarks, RegLand is able to guarantee an MSE of 0.016 whereas RegLand-Sp offers an MSE of less than 0.02 ; MLKR is only able to guarantee an MSE rate of 0.04 for this dataset. In Table 1a, we compare accuracies of the two algorithms when given 50 landmark points with those of KR for the Sigmoid and Manhattan kernels. We find that in all cases, RegLand-Sp gives superior accuracies than KR. Moreover, the Manhattan kernel seems to match or outperform the Sigmoid kernel on all the datasets.

Ordinal Regression: Here, we compare our method with the baseline KR method on benchmark datasets. As mentioned in Section 3.3, our method uses the EXC formulation of [16] along with landmarking scheme given in Algorithm 1. We implemented a gradient descent-based solver (ORLand) to solve the primal formulation of EXC and used fixed equi-spaced thresholds instead of learning them as suggested by [16]. Of the six datasets considered here, the two `Wine` datasets are ordinal regression datasets where the quality of the wine is to be predicted on a scale from 1 to 10. The remaining four datasets are regression datasets whose labels were subjected to equi-frequency binning to obtain ordinal regression datasets [16]. We measured the average absolute error (AAE) for each method. Figure 1b compares ORLand with KR as the number of landmarks increases. Table 1b compares accuracies of ORLand for 50 landmark points with those of KR for Sigmoid and Manhattan kernels. In almost all cases, ORLand gives a much better performance than KR. The Sigmoid kernel seems to outperform the Manhattan kernel on a couple of datasets.

We refer the reader to Appendix G for additional experimental results.

5 Conclusion

In this work we considered the general problem of supervised learning using non-PSD similarity functions. We provided a goodness criterion for similarity functions w.r.t. various learning tasks. This allowed us to construct efficient learning algorithms with provable generalization error bounds. At the same time, we were able to show, for each learning task, that our criterion is not too restrictive in that it admits all good PSD kernels. We then focused on the problem of identifying influential landmarks with the aim of learning sparse predictors. We presented a model that formalized the intuition that typically only a small fraction of landmarks is influential for a given learning problem. We adapted existing sparse vector recovery algorithms within our model to learn provably sparse predictors with bounded generalization error. Finally, we empirically evaluated our learning algorithms on benchmark regression and ordinal regression tasks. In all cases, our learning methods, especially the sparse recovery algorithm, consistently outperformed the kernel regression baseline.

An interesting direction for future research would be learning good similarity functions á la metric learning or kernel learning. It would also be interesting to conduct large scale experiments on real-world data such as social networks that naturally capture the notion of similarity amongst nodes.

Acknowledgments

P. K. is supported by a Microsoft Research India Ph.D. fellowship award. Part of this work was done while P. K. was an intern at Microsoft Research Labs India, Bangalore.

References

[1] Bernhard Schölkopf and Alex J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[2] Bernard Haasdonk. Feature Space Interpretation of SVMs with Indefinite Kernels. *IEEE Transactions on Pattern Analysis and Machince Intelligence*, 27(4):482–492, 2005.

[3] Cheng Soon Ong, Xavier Mary, Stéphane Canu, and Alexander J. Smola. Learning with non-positive Kernels. In *21st Annual International Conference on Machine Learning*, 2004.

[4] Yihua Chen, Maya R. Gupta, and Benjamin Recht. Learning Kernels from Indefinite Similarities. In *26th Annual International Conference on Machine Learning*, pages 145–152, 2009.

[5] Ronny Luss and Alexandre d'Aspremont. Support Vector Machine Classification with Indefinite Kernels. In *21st Annual Conference on Neural Information Processing Systems*, 2007.

[6] Maria-Florina Balcan and Avrim Blum. On a Theory of Learning with Similarity Functions. In *23rd Annual International Conference on Machine Learning*, pages 73–80, 2006.

[7] Liwei Wang, Cheng Yang, and Jufu Feng. On Learning with Dissimilarity Functions. In *24th Annual International Conference on Machine Learning*, pages 991–998, 2007.

[8] Purushottam Kar and Prateek Jain. Similarity-based Learning via Data Driven Embeddings. In *25th Annual Conference on Neural Information Processing Systems*, 2011.

[9] Nathan Srebro. How Good Is a Kernel When Used as a Similarity Measure? In *20th Annual Conference on Computational Learning Theory*, pages 323–335, 2007.

[10] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading Accuracy for Sparsity in Optimization Problems with Sparsity Constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

[11] Nathan Srebro Shai Ben-David, Ali Rahimi. Generalization Bounds for Indefinite Kernel Machines. In *NIPS 2008 Workshop: New Challenges in Theoretical Machine Learning*, 2008.

[12] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based Classification: Concepts and Algorithms. *Journal of Machine Learning Research*, 10:747–776, 2009.

[13] Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the Complexity of Linear Prediction : Risk Bounds, Margin Bounds, and Regularization. In *22nd Annual Conference on Neural Information Processing Systems*, 2008.

[14] Chia-Hua Ho and Chih-Jen Lin. Large-scale Linear Support Vector Regression. `http://www.csie.ntu.edu.tw/~cjlin/papers/linear-svr.pdf`, retrieved on May 18, 2012, 2012.

[15] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. Improved Guarantees for Learning via Similarity Functions. In *21st Annual Conference on Computational Learning Theory*, pages 287–298, 2008.

[16] Wei Chu and S. Sathiya Keerthi. Support Vector Ordinal Regression. *Neural Computation*, 19(3):792–815, 2007.

[17] Shivani Agarwal. Generalization Bounds for Some Ordinal Regression Algorithms. In *19th International Conference on Algorithmic Learning Theory*, pages 7–21, 2008.

[18] A. Frank and Arthur Asuncion. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml`, 2010. University of California, Irvine, School of Information and Computer Sciences.

[19] StatLib Dataset Repository. `http://lib.stat.cmu.edu/datasets/`. Carnegie Mellon University.

[20] Delve Dataset Repository. `http://www.cs.toronto.edu/~delve/data/datasets.html`. University of Toronto.

[21] Kilian Q. Weinberger and Gerald Tesauro. Metric Learning for Kernel Regression. In *11th International Conference on Artificial Intelligence and Statistics*, pages 612–619, 2007.

[22] Tong Zhang. Covering Number Bounds of Certain Regularized Linear Function Classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

[23] Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities : Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[24] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A Generalized Representer Theorem. In *14th Annual Conference on Computational Learning Theory*, pages 416–426, 2001.

[25] Colin McDiarmid. On the Method of Bounded Differences. *Surveys in Combinatorics*, 141:148–188, 1989.

[26] Amiran Ambroladze and John Shawe-Taylor. Complexity of Pattern Classes and Lipschitz Property. In *15th International Conference on Algorithmic Learning Theory*, pages 181–193, 2004.

[27] Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On NDCG Consistency of Listwise Ranking Methods. In *14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR Conference and Workshop Proceedings*, pages 618–626, 2011.

[28] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, 2000.

[29] Shivani Agarwal and Partha Niyogi. Generalization Bounds for Ranking Algorithms via Algorithmic Stability. *Journal of Machine Learning Research*, 10:441–474, 2009.

Supplementary Material

Throughout this document, theorems and lemmata that were not originally proven as a part of this work cite, as a part of their statement, the work that originally presented the proof.

Appendix A Proofs of supplementary theorems

In this section we give proofs for certain generic results that would be used in the utility and admissibility proofs. The first result, given as Lemma 15, allows us to analyze the landmarking step (Step 1 of Algorithm 1) and allows us to reduce the learning problem to that of learning a linear predictor over the landmarked space. The second result, given as Lemma 16, gives us a succinct re-statement of generalization error bounds proven in [13] that would be used in proving utility bounds. The third result, given as Lemma 17, is a technical result that helps us prove admissibility bounds for our goodness definitions.

Lemma 15 (Landmarking approximation guarantee [8])**.** *Given a similarity function K over a domain $\mathcal{X}$ and a bounded function of the form $f(x) = \mathop{\mathbb{E}}_{\mathbf{x}' \sim \mathcal{D}} [\![w(\mathbf{x}') K(\mathbf{x}, \mathbf{x}')]\!]$ for some bounded weight function $w : \mathcal{X} \to \{-B, B\}$, for every $\epsilon, \delta > 0$ there exists a randomized map $\Psi : \mathcal{X} \to \mathbb{R}^d$ for $d = d(\epsilon, \delta)$ such that with probability at least $1 - \delta$, there exists a linear operator $\tilde{f}$ over $\mathbb{R}^d$ such that $\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\left| \tilde{f}(\Psi(\mathbf{x})) - f(\mathbf{x}) \right| \right]\!\!\right] \leq \epsilon$.*

Proof. This result essentially allows us to project the learning problem into a Euclidean space where one can show, for the various learning problems considered here, that existing large margin techniques are applicable to solve the original problem. The result appeared in [8] and is presented here for completeness.

Sample d *landmark points* $\mathcal{L} = \{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$ from $\mathcal{D}$ and construct the map $\Psi_{\mathcal{L}} : \mathbf{x} \mapsto \frac{1}{\sqrt{d}} (K(\mathbf{x}, \mathbf{x}_1), \ldots, K(\mathbf{x}, \mathbf{x}_d))$ and consider the linear operator $\tilde{f}$ over $\mathbb{R}^d$ defined as follows (in the following, we shall always omit the subscript $\mathcal{L}$ for clarity):

$$\tilde{f} : \mathbf{x} \mapsto \frac{1}{d} \sum_{i=1}^{d} w(\mathbf{x}_i) K(\mathbf{x}, \mathbf{x}_i) = \langle \tilde{\mathbf{w}}, \Psi(\mathbf{x}) \rangle$$

for $\mathbf{w} = \frac{1}{\sqrt{d}} (w(\mathbf{x}_1), \ldots, w(\mathbf{x}_d)) \in \mathbb{R}^d$. A standard Hoeffding-style argument shows that for $d = \mathcal{O}\left(\frac{B^2}{\epsilon^2} \log \frac{1}{\delta^2} \right) = \mathcal{O}\left(\frac{B^2}{\epsilon^2} \log \frac{1}{\delta} \right)$, $\tilde{f}$ gives a point wise approximation to f, i.e. for all $\mathbf{x} \in \mathcal{X}$, with probability greater than $1 - \delta^2$, we have $\left| \tilde{f}(\Psi(\mathbf{x})) - f(\mathbf{x}) \right| < \epsilon$.

Now call the event $\textsf{BAD-APPROX}(\mathbf{x}) := \left| \tilde{f}(\Psi(\mathbf{x})) - f(\mathbf{x}) \right| > \epsilon$. Thus we have for all $\mathbf{x} \in \mathcal{X}$, $\mathop{\mathbb{P}}_{\tilde{f}} [\textsf{BAD-APPROX}(\mathbf{x})] = \mathop{\mathbb{E}}_{\tilde{f}} [\![1_{\textsf{BAD-APPROX}(\mathbf{x})}]\!] < \delta^2$ (here the probabilities are being taken over

the construction of $\tilde{f}$ i.e. the choice of the landmark points). Taking expectations over the entire domain, applying Fubini's theorem to switch expectations and applying Markov's inequality we get

$$\mathop{\mathbb{P}}_{\tilde{f}}\left[\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}\left[\textsf{BAD-APPROX}\left(\mathbf{x}\right)\right] > \delta\right] < \delta$$

Thus with confidence $1-\delta$ we have $\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}\left[\textsf{BAD-APPROX}\left(\mathbf{x}\right)\right] < \delta$ and thus $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\left|\tilde{f}(\Psi(\mathbf{x})) - f(\mathbf{x})\right|\right]\!\right] < (1-\delta)\epsilon + 2B\delta$ since $\sup_{\mathbf{x}\in\mathcal{X}}\left|\tilde{f}(\Psi(\mathbf{x}))\right| = \sup_{\mathbf{x}\in\mathcal{X}}|f(\mathbf{x})| = B$. For $\delta < \frac{\epsilon}{B}$ we get $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\left|\tilde{f}(\Psi(\mathbf{x})) - f(\mathbf{x})\right|\right]\!\right] < 2\epsilon$. □

Lemma 16 (Risk bounds for linear predictors [13])**.** *Consider a real-valued prediction problem y over a domain $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq C_X\}$ and a linear learning model $\mathcal{F} : \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\|_2 \leq C_W\}$ under some fixed loss function $\ell(\cdot,\cdot)$ that is C_L-Lipschitz in its second argument. For any $f \in \mathcal{F}$, let $\mathcal{L}_f = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell(f(\mathbf{x}), y(\mathbf{x}))\right]\!\right]$ and $\hat{\mathcal{L}}_f^n$ be the empirical loss on a set of n i.i.d. chosen points. Then we have, with probability greater than $(1-\delta)$,*

$$\sup_{f\in\mathcal{F}}\left(\mathcal{L}_f - \hat{\mathcal{L}}_f^n\right) \leq 3C_L C_X C_W\sqrt{\frac{\log(1/\delta)}{n}}$$

Proof. There exist a few results that provide a unified analysis for the generalization properties of linear predictors [13, 22]. However we use the heavy hammer of Rademacher average based analysis since it provides sharper bounds than covering number based analyses.

The result follows from imposing a squared L_2 regularization on the $\mathbf{w}$ vectors. Since the squared L_2 function is 2-strongly convex with respect to the L_2 norm, using [13, Theorem 1], we get a bound on the Rademacher complexity of the function class $\mathcal{F}$ as $\mathcal{R}_n(\mathcal{F}) \leq C_X C_W\sqrt{\frac{1}{n}}$. Next, using the Lipschitz properties of the loss function, a result from [23] allows us to bound the excess error by $2C_L\mathcal{R}_n(\mathcal{F}) + C_L C_X C_W\sqrt{\frac{\log(1/\delta)}{2n}}$. The result then follows from simple manipulations. □

Lemma 17 (Admissible weight functions for PSD kernels [9])**.** *Consider a PSD kernel that is (ϵ_0,γ)-good for a learning problem with respect to some convex loss function ℓ_K. Then there exists a vector $\mathbf{W}' \in \mathcal{H}_K$ and a bounded weight function $w : \mathcal{X} \to \mathbb{R}$ such that $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_K\left(\langle \mathbf{W}', \Phi_K(\mathbf{x})\rangle, y(\mathbf{x})\right)\right]\!\right] \leq \epsilon_0 + \frac{1}{2C\gamma^2}$ for some arbitrary positive constant C and for all $\mathbf{x} \in \mathcal{X}$, we have $\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}\left[\!\left[w(\mathbf{x}')K(\mathbf{x},\mathbf{x}')\right]\!\right] = \langle \mathbf{W}', \Phi_K(\mathbf{x})\rangle$.*

Proof. Note that the (ϵ_0,γ)-goodness of K guarantees the existence of a weight vector $\mathbf{W}^* \in \mathcal{H}_K$ with small loss at large margin. Thus $\mathbf{W}'$ acts as a proxy for $\mathbf{W}^*$ providing bounded loss at unit margin but with the additional property of being functionally equivalent to a bounded weighted average of the kernel values as required by the definition of a good similarity function. This will help us prove admissibility results for our similarity learning models.

We start by proving the theorem for a discrete distribution - the generalization to non-discrete distributions will follow by using variational optimization techniques as discussed in [9]. Consider a discrete learning problem with $\mathcal{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, corresponding distribution $\mathcal{D} = \{p_1,\ldots,p_n\}$ and target $y = \{y_1,\ldots,y_n\}$ such that $\sum p_i = 1$. Set up the following regularized ERM problem (albeit on the entire domain):

$$\min_{\mathbf{W}\in\mathcal{H}_K} \quad \frac{1}{2}\|\mathbf{W}\|_{\mathcal{H}_K}^2 + C\sum_{i=1}^n p_i\ell_K\left(\langle \mathbf{W}, \Phi_K(\mathbf{x}_i)\rangle, y_i\right)$$

Let $\mathbf{W}'$ be the weight vector corresponding to the optima of the above problem. By the Representer Theorem (for example [24]), we can choose $\mathbf{W}' = \sum \alpha_i\Phi_K(\mathbf{x}_i)$ for some bounded α_i (the exact

bounds on α_i are problem specific). By (ϵ_0, γ)-goodness of K we have

$$\begin{aligned}
\frac{1}{2}\left\|\mathbf{W}'\right\|_{\mathcal{H}_K}^2 + C\sum_{i=1}^{n} p_i \ell_K\left(\left\langle \mathbf{W}', \Phi_K(\mathbf{x}_i)\right\rangle, y_i\right) &\leq \frac{1}{2}\left\|\frac{1}{\gamma}\mathbf{W}^*\right\|_{\mathcal{H}_K}^2 + C\sum_{i=1}^{n} p_i \ell_K\left(\frac{\left\langle \mathbf{W}^*, \Phi_K(\mathbf{x}_i)\right\rangle}{\gamma}, y_i\right) \\
&= \frac{1}{2\gamma^2} + C\cdot \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_K\left(\frac{\left\langle \mathbf{W}^*, \Phi_K(\mathbf{x})\right\rangle}{\gamma}, y(\mathbf{x})\right)\right]\!\!\right] \\
&\leq \frac{1}{2\gamma^2} + C\epsilon_0
\end{aligned}$$

Thus we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_K\left(\left\langle \mathbf{W}', \Phi_K(\mathbf{x})\right\rangle, y(\mathbf{x})\right)\right]\!\!\right] &\leq \frac{1}{2C}\left\|\mathbf{W}'\right\|_{\mathcal{H}_K}^2 + \sum_{i=1}^{n} p_i \ell_K\left(\left\langle \mathbf{W}', \Phi_K(\mathbf{x}_i)\right\rangle, y_i\right) \\
&\leq \epsilon_0 + \frac{1}{2C\gamma^2}
\end{aligned}$$

which proves the first part of the claim. For the second part, set up a weight function $w_i = \frac{\alpha_i}{p_i}$. Then, for any $\mathbf{x} \in \mathcal{X}$ we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}\left[\!\!\left[w(\mathbf{x}')K(\mathbf{x}, \mathbf{x}')\right]\!\!\right] &= \sum_{i=1}^{n} p_i w_i K(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^{n} p_i \frac{\alpha_i}{p_i} K(\mathbf{x}, \mathbf{x}_i) \\
&= \sum_{i=1}^{n} \alpha_i \left\langle \Phi_K(\mathbf{x}), \Phi_K(\mathbf{x}_i)\right\rangle = \left\langle \mathbf{W}', \Phi_K(\mathbf{x})\right\rangle
\end{aligned}$$

The weight function is bounded since the α_i are bounded and, this being a discrete learning problem, cannot have vanishing probability masses p_i (actually, in the cases we shall consider, the α_i will itself contain a p_i term that will subsequently get cancelled). For non-discrete cases, variational techniques give us similar results. □

Appendix B Justifying Double-dipping

All our analyses (as well as the analyses presented in [6, 7, 8]) use some data as landmark points and then require a fresh batch of training points to learn a classifier on the landmarked space. In practice, however, it might be useful to reuse training data to act as landmark points as well. This is especially true of [7, 8] who require labeled landmarks. We give below, generalization bounds for similarity-based learning algorithms that indulge in such "double dipping". The argument uses a technique outlined in [11] and falls within the Rademacher-average based uniform convergence guarantees used elsewhere in the paper. We present a generic argument that, in a manner similar to Lemma 16, can be specialized to the various learning problems considered in this paper.

To make the presentation easier we set up some notation. For any predictor f, let $\mathcal{L}_f = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell(f(\mathbf{x}), y(\mathbf{x}))\right]\!\!\right]$ and for any training set S of size n, let $\hat{\mathcal{L}}_f^S = \frac{1}{n}\sum_{\mathbf{x}_i\in S}\ell(f(\mathbf{x}_i), y(\mathbf{x}_i))$. For any landmark set $S = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, we let $\Psi_S : \mathbf{x} \mapsto (K(\mathbf{x}, \mathbf{x}_1), \ldots, K(\mathbf{x}, \mathbf{x}_n))$. For any weight vector $\mathbf{w} \in \mathbb{R}^n, \|\mathbf{w}\|_\infty \leq B$ in the landmarked space, denote the predictor $f_{(S,\mathbf{w})} := \frac{1}{n}\left\langle \mathbf{w}, \Psi_S(\mathbf{x})\right\rangle = \mathbf{x} \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_i K(\mathbf{x}, \mathbf{x}_i)$. Also let $\mathcal{F}_S := \left\{\mathbf{x} \mapsto \frac{1}{n}\left\langle \mathbf{w}, \Psi_S(\mathbf{x})\right\rangle\right\} = \left\{f_{(S,\mathbf{w})} : \mathbf{w} \in \mathbb{R}^n, \|\mathbf{w}\|_\infty \leq B\right\}$.

We note that the embedding defined above is "stable" in the sense that changing a single landmark does not change the embedding too much with respect to bounded predictors. More formally, for any set of n points $S = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, define $g(S) := \sup_{f\in\mathcal{F}_S}\left\{\mathcal{L}_f - \hat{\mathcal{L}}_f^S\right\}$. Let S^i be another set of n points that (arbitrarily) differs from S just at the i^{th} point and coincides with S on the rest. Then we have, for any fixed $\mathbf{w}$ of bounded L_∞ norm (i.e. $\|\mathbf{w}\|_\infty \leq B$) and bounded similarity function (i.e.

$K(\mathbf{x}, \mathbf{y}) \leq 1$),

$$\begin{aligned}
\sup_{\mathbf{x}} \left\{ \left| f_{(S,\mathbf{w})}(\mathbf{x}) - f_{(S^i,\mathbf{w})}(\mathbf{x}) \right| \right\} &= \sup_{\mathbf{x}} \left\{ \left| \frac{1}{n} \sum_{j=1}^{n} \mathbf{w}_j K(\mathbf{x}, \mathbf{x}_j) - \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_j K(\mathbf{x}, \mathbf{x}'_j) \right| \right\} \\
&= \sup_{\mathbf{x}} \left\{ \left| \frac{1}{n} \mathbf{w}_i \left(K(\mathbf{x}, \mathbf{x}_i) - K(\mathbf{x}, \mathbf{x}'_i) \right) \right| \right\} \\
&\leq \frac{2B}{n}
\end{aligned}$$

Note that, although [8] uses pairs of labeled points to define the embedding, the following argument can easily be extended to incorporate this since the embedding is identical to the embedding Ψ_S described above with respect to being "stable". In fact this analysis holds for any stable embedding defined using training points.

Our argument proceeds by showing that with high probability (over choice of the set S) we have

$$\sup_{\mathbf{w}} \left\{ \mathcal{L}_{f_{(S,\mathbf{w})}} - \hat{\mathcal{L}}^S_{f_{(S,\mathbf{w})}} \right\} \leq \epsilon$$

By the definition of $\mathcal{F}_S$, the above requirement translates to showing that with high probability,

$$\sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}^S_f \right\} \leq \epsilon$$

which highlights the fact that we are dealing with a problem of sample dependent hypothesis spaces[1]. Note that this exactly captures the double dipping procedure of reusing training points as landmark points. Such a result would be useful as follows: using Lemma 15 and task specific guarantees (outlined in detail in the subsequent sections), we have, with high probability, the existence of a good predictor in the landmarked space of a randomly chosen landmark set S i.e. with very high probability over choice of S, we have $\inf_{f \in \mathcal{F}_S} \{\mathcal{L}_f\} \leq \epsilon_0$. Let this be achieved by the predictor f^*. Using the uniform convergence guarantee above we get $\hat{\mathcal{L}}^S_{f^*} \leq \epsilon_0 + \epsilon$ (with some loss of confidence due to application of a union bound).

Now consider the predictor $\hat{f} := \inf_{f \in \mathcal{F}_S} \left\{ \hat{\mathcal{L}}^S_f \right\}$. Clearly $\hat{\mathcal{L}}^S_{\hat{f}} \leq \hat{\mathcal{L}}^S_{f^*} \leq \epsilon_0 + \epsilon$. Invoking the uniform convergence bound yet again shows us that

$$\mathcal{L}_{\hat{f}} \leq \hat{\mathcal{L}}^S_{\hat{f}} + \sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}^S_f \right\} \leq \epsilon_0 + 2\epsilon$$

Note that we incur some more loss of confidence due to another application of the union bound. This tells us that with high probability, a predictor learned by choosing a random landmark set and training on the landmark set itself would yield a good predictor.

We will proceed via a vanilla uniform convergence argument involving symmetrization and an application of the McDiarmid's inequality (stated below). However, proving the stability prerequisite for the application of the McDiarmid's inequality shall require use of the stability of both the predictor $f_{(S,\mathbf{w})}$ as well as the embedding Ψ_S. Let the loss function ℓ be C_L-Lipschitz in its first argument.

Theorem 18 (McDiarmid's inequality [25])**.** *Let $X_1, \ldots, X_n$ be independent random variables taking values in some set $\mathcal{X}$. Further, let $f : \mathcal{X}^n \to \mathbb{R}$ be a function of n variables that satisfies, for all $i \in [n]$ and all $x_1, \ldots, x_n, x'_i \in \mathcal{X}$,*

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x'_i, \ldots, x_n)| \leq c_i$$

then for all $\epsilon > 0$, we have

$$\mathbb{P}\left[f - \mathbb{E}\left[\!\left[f \right]\!\right] > \epsilon \right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2} \right)$$

[1]We were not able to find any written manuscript detailing the argument of [11]. However the argument itself is fairly generic in allowing one to prove generalization bounds for sample dependent hypothesis spaces.

We shall invoke the McDiarmid's inequality on the function $g(S) := \sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}_f^S \right\}$ with $S = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ being the random variables in question. To do so we first prove the stability of the function $g(S)$ with respect to its variables and then bound the value of $\mathop{\mathbb{E}}_S [\![g(S)]\!]$.

Theorem 19. *For any S, S^i, we have $\left| g(S) - g(S^i) \right| \leq \frac{6BC_L}{n}$.*

Proof. We have

$$\begin{aligned} g(S) &= \sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}_f^S \right\} \\ &= \sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}_f^S - \hat{\mathcal{L}}_f^{S^i} + \hat{\mathcal{L}}_f^{S^i} \right\} \\ &\leq \sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}_f^{S^i} \right\} + \sup_{f \in \mathcal{F}_S} \left\{ \hat{\mathcal{L}}_f^S - \hat{\mathcal{L}}_f^{S^i} \right\} \\ &\leq \sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}_f^{S^i} \right\} + \frac{2BC_L}{n} \end{aligned}$$

where in the fourth step we have used the fact that the loss function is Lipschitz and the embedding function Ψ_S is bounded. We also have

$$\begin{aligned} \sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}_f^{S^i} \right\} &= \sup_{\mathbf{w}} \left\{ \mathcal{L}_{f_{(S,\mathbf{w})}} - \hat{\mathcal{L}}^{S^i}_{f_{(S,\mathbf{w})}} \right\} \\ &= \sup_{\mathbf{w}} \left\{ \mathcal{L}_{f_{(S,\mathbf{w})}} - \mathcal{L}_{f_{(S^i,\mathbf{w})}} + \mathcal{L}_{f_{(S^i,\mathbf{w})}} - \hat{\mathcal{L}}^{S^i}_{f_{(S^i,\mathbf{w})}} + \hat{\mathcal{L}}^{S^i}_{f_{(S^i,\mathbf{w})}} - \hat{\mathcal{L}}^{S^i}_{f_{(S,\mathbf{w})}} \right\} \\ &\leq \sup_{\mathbf{w}} \left\{ \mathcal{L}_{f_{(S^i,\mathbf{w})}} - \hat{\mathcal{L}}^{S^i}_{f_{(S^i,\mathbf{w})}} \right\} + \sup_{\mathbf{w}} \left\{ \mathcal{L}_{f_{(S,\mathbf{w})}} - \mathcal{L}_{f_{(S^i,\mathbf{w})}} \right\} \\ &\quad + \sup_{\mathbf{w}} \left\{ \hat{\mathcal{L}}^{S^i}_{f_{(S^i,\mathbf{w})}} - \hat{\mathcal{L}}^{S^i}_{f_{(S,\mathbf{w})}} \right\} \\ &\leq \sup_{\mathbf{w}} \left\{ \mathcal{L}_{f_{(S^i,\mathbf{w})}} - \hat{\mathcal{L}}^{S^i}_{f_{(S^i,\mathbf{w})}} \right\} + \frac{2BC_L}{n} + \frac{2BC_L}{n} \\ &= \sup_{f \in \mathcal{F}_{S^i}} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}_f^{S^i} \right\} + \frac{4BC_L}{n} \\ &= g(S^i) + \frac{4BC_L}{n} \end{aligned}$$

where in the fourth step we have used the stability of the embedding function and that the loss function is C_L-Lipschitz in its first argument so that for all $\mathbf{x}$ we have $\left| \ell\left(f_{(S,\mathbf{w})}(\mathbf{x}), y(\mathbf{x}) \right) - \ell\left(f_{(S^i,\mathbf{w})}(\mathbf{x}), y(\mathbf{x}) \right) \right| \leq \frac{2BC_L}{n}$ which holds in expectation over any (empirical) distribution as well. Putting the two inequalities together gives us $g(S) \leq g(S^i) + \frac{6BC_L}{n}$. Similarly we also have $g(S^i) \leq g(S) + \frac{6BC_L}{n}$ which gives us the result. □

We now have that the function $g(S)$ is $\mathcal{O}\left(\frac{1}{n}\right)$-stable with respect to each of its inputs. We now move on to bound its expectation. For any function class $\mathcal{F}$ we define its *empirical Rademacher average* as follows

$$\hat{\mathcal{R}}_n(\mathcal{F}) := \mathop{\mathbb{E}}_{\sigma} \left[\!\!\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{\mathbf{x}_i \in S} \sigma_i f(\mathbf{x}_i) \right\} \middle| S \right]\!\!\right]$$

Also let $\mathcal{F} := \{ \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_2 \leq B \}$ and $\mathcal{X} := \{ \mathbf{x} : \|\mathbf{x}\|_2 \leq 1 \}$.

Theorem 20. $\mathop{\mathbb{E}}_S \left[\!\!\left[\sup_{f \in \mathcal{F}_S} \left\{ \mathcal{L}_f - \hat{\mathcal{L}}_f^S \right\} \right]\!\!\right] \leq 2BC_L \sqrt{\frac{1}{n}}$

Proof. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}\left[\!\!\left[\sup_{f\in\mathcal{F}_S}\left\{\mathcal{L}_f-\hat{\mathcal{L}}_f^S\right\}\right]\!\!\right] &= \mathop{\mathbb{E}}_{S}\left[\!\!\left[\sup_{f\in\mathcal{F}_S}\left\{\mathop{\mathbb{E}}_{S'}\left[\!\!\left[\hat{\mathcal{L}}_f^{S'}\right]\!\!\right]-\hat{\mathcal{L}}_f^S\right\}\right]\!\!\right] \\
&\leq \mathop{\mathbb{E}}_{S,S'}\left[\!\!\left[\sup_{f\in\mathcal{F}_S}\left\{\hat{\mathcal{L}}_f^{S'}-\hat{\mathcal{L}}_f^S\right\}\right]\!\!\right] \\
&\leq \mathop{\mathbb{E}}_{S,S'}\left[\!\!\left[\sup_{f\in\mathcal{F}_{S\cup S'}}\left\{\hat{\mathcal{L}}_f^{S'}-\hat{\mathcal{L}}_f^S\right\}\right]\!\!\right] \\
&= \mathop{\mathbb{E}}_{S,S',\sigma}\left[\!\!\left[\sup_{f\in\mathcal{F}_{S\cup S'}}\left\{\frac{1}{n}\sum_{\mathbf{x}_i\in S,\mathbf{x}_i'\in S'}\sigma_i\left(\ell(f(\mathbf{x}_i'),y(\mathbf{x}_i'))-\ell(f(\mathbf{x}_i),y(\mathbf{x}_i))\right)\right\}\right]\!\!\right] \\
&\leq 2\mathop{\mathbb{E}}_{S,S',\sigma}\left[\!\!\left[\sup_{f\in\mathcal{F}_{S\cup S'}}\left\{\frac{1}{n}\sum_{\mathbf{x}_i\in S}\sigma_i\ell(f(\mathbf{x}_i),y(\mathbf{x}_i))\right\}\right]\!\!\right] \\
&= 2\mathop{\mathbb{E}}_{S,S',\sigma}\left[\!\!\left[\sup_{\mathbf{w}}\left\{\frac{1}{n}\sum_{\mathbf{x}_i\in S}\sigma_i\ell(f_{(S\cup S',\mathbf{w})}(\mathbf{x}_i),y(\mathbf{x}_i))\right\}\right]\!\!\right] \\
&\leq 2\mathop{\mathbb{E}}_{S,S',\sigma}\left[\!\!\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{\mathbf{x}_i\in S}\sigma_i\ell(f(\mathbf{x}_i),y(\mathbf{x}_i))\right\}\right]\!\!\right] \\
&= 2\mathop{\mathbb{E}}_{S}\left[\!\!\left[\hat{\mathcal{R}}_n(\ell\circ\mathcal{F})\right]\!\!\right]\leq 2C_L\mathop{\mathbb{E}}_{S}\left[\!\!\left[\hat{\mathcal{R}}_n(\mathcal{F})\right]\!\!\right]\leq 2BC_L\sqrt{\frac{1}{n}}
\end{aligned}$$

where in the third step we have used the fact that $\mathcal{F}_S \supseteq \mathcal{F}_{S'}$ if $S \supseteq S'$ (this is the monotonicity requirement in [11]). Note that this is essential to introduce symmetry so that Rademacher variables can be introduced in the next (symmetrization) step. In the seventh step, we have used the fact that for every S such that $|S| = n$ and $\mathbf{w} \in \mathbb{R}^n$ such that $\|\mathbf{w}\|_\infty \leq B$, there exists a function $f \in \mathcal{F}$ such that for all $\mathbf{x}$, there exists a $\mathbf{x}' \in \mathcal{X}$ such that $f_{(S,\mathbf{w})}(\mathbf{x}) = f(\mathbf{x}')$. In the last step we have used a result from [26] which allows calculation of Rademacher averages for composition classes and an intermediate result from the proof of Lemma 16 which gives us Rademacher averages for the function class $\mathcal{F}$. □

Thus, by an application of McDiarmid's inequality we have, with probability $(1-\delta)$ over choice of the landmark (training) set,

$$\sup_{f\in\mathcal{F}_S}\left\{\mathcal{L}_f-\hat{\mathcal{L}}_f^S\right\}\leq\mathbb{E}\left[\!\!\left[\sup_{f\in\mathcal{F}_S}\left\{\mathcal{L}_f-\hat{\mathcal{L}}_f^S\right\}\right]\!\!\right]+6BC_L\sqrt{\frac{\log 1/\delta}{2n}}\leq 4BC_L\sqrt{\frac{\log 1/\delta}{n}}$$

which concludes our argument justifying double dipping.

Appendix C Regression with Similarity Functions

In this section we give proofs of utility and admissibility results for our similarity based learning model for real-valued regression tasks.

C.1 Proof of Theorem 5

First of all, we use Lemma 15 to project onto a d dimensional space where there exists a linear predictor $\tilde{f} : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ such that $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\left|\tilde{f}\left(\Psi\left(\mathbf{x}\right)\right)-f(\mathbf{x})\right|\right]\!\!\right] \leq 2\epsilon_1$. Note that $\|\mathbf{w}\|_2 \leq B$ and $\sup_{\mathbf{x}\in\mathcal{X}}\{\|\Psi(\mathbf{x})\|\} \leq 1$ by construction. We will now show that $\tilde{f}$ has bounded ϵ-insensitive loss.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_\epsilon\left(\tilde{f}\left(\Psi\left(\mathbf{x}\right)\right),y(\mathbf{x})\right)\right]\!\!\right] &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_\epsilon\left(f(\mathbf{x}),y(\mathbf{x})\right)\right]\!\!\right] + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_\epsilon\left(\tilde{f}\left(\Psi\left(\mathbf{x}\right)\right),y(\mathbf{x})\right)-\ell_\epsilon\left(f(\mathbf{x}),y(\mathbf{x})\right)\right]\!\!\right] \\
&\leq \epsilon_0 + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_\epsilon\left(\tilde{f}\left(\Psi\left(\mathbf{x}\right)\right),y(\mathbf{x})\right)-\ell_\epsilon\left(f(\mathbf{x}),y(\mathbf{x})\right)\right]\!\!\right] \\
&\leq \epsilon_0 + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\left|\tilde{f}\left(\Psi(\mathbf{x})\right)-f(\mathbf{x})\right|\right]\!\!\right] \\
&\leq \epsilon_0 + 2\epsilon_1
\end{aligned}$$

where in the second step we have used the goodness properties of K, in the third step we used the fact that the ϵ-insensitive loss function is 1-Lipschitz in its first argument. Note that $\|\mathbf{w}\| \approx \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[w^2(\mathbf{x})\right]\!\!\right]$ with high probability and if $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[w^2(\mathbf{x})\right]\!\!\right] \ll B$ then we get a much better bound on the norm of $\mathbf{w}$. The excess loss incurred due to this landmarking step is, with probability $1-\delta$, at most $32B\sqrt{\frac{\log(1/\delta)}{d}}$.

Now consider the following regularized ERM problem on n i.i.d. sample points:

$$\hat{\mathbf{w}} = \mathop{\arg\min}_{\mathbf{w}:\|\mathbf{w}\|_2\leq B} \frac{1}{n}\sum_{i=1}^{n} \ell_\epsilon\left(\left\langle\mathbf{w},\Psi(\mathbf{x}_i)\right\rangle,y(\mathbf{x}_i)\right)$$

The final output of our learning algorithm shall be $\mathbf{x}\mapsto\langle\hat{\mathbf{w}},\Psi(\mathbf{x})\rangle$. Here we have $C_X=1$, $C_L=1$ since $\ell_\epsilon(\cdot)$ is 1-Lipschitz and $C_W=B$. Thus by Lemma 16, we get that the excess loss incurred due to this regularized ERM step is at most $3B\sqrt{\frac{\log 1/\delta}{n}}$.

Since the ϵ-insensitive loss is related to the absolute error by $|x|\leq\ell_\epsilon(x)+\epsilon$ we have the total error (with respect to absolute loss) being incurred by our predictor to be, with probability at least $1-2\delta$, at most

$$\epsilon_0 + 32B\sqrt{\frac{\log(1/\delta)}{d}} + 3B\sqrt{\frac{\log 1/\delta}{n}} + \epsilon$$

Taking $d=\mathcal{O}\left(\frac{B^2}{\epsilon_1^2}\log\frac{1}{\delta}\right)$ unlabeled landmarks and $n=\mathcal{O}\left(\frac{B^2}{\epsilon_1^2}\log\frac{1}{\delta}\right)$ labeled training points gives us our desired result.

C.2 Proof of Theorem 6

We prove the two parts of the result separately.

Part 1: Admissibility: Using Lemma 17 it is possible to obtain a vector $\mathbf{W}'=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\Phi_K(\mathbf{x}_i)\in\mathcal{H}_K$ with small loss such that $0\leq\alpha_i,\alpha_i^*\leq p_iC$ and $\alpha_i\alpha_i^*=0$ (these inequalities are a consequence of applying the KKT conditions). This allows us to construct as weight function $w_i=\frac{\alpha_i-\alpha_i^*}{p_i}$ such that $|w_i|\leq C$ and $\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}\left[\!\!\left[w(\mathbf{x}')K(\mathbf{x},\mathbf{x}')\right]\!\!\right]=\langle\mathbf{W}',\Phi_K(\mathbf{x})\rangle$ for all $\mathbf{x}\in\mathcal{X}$.

Thus we have $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_\epsilon\left(\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}\left[\!\!\left[w(\mathbf{x}')K(\mathbf{x},\mathbf{x}')\right]\!\!\right],y(\mathbf{x})\right)\right]\!\!\right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\ell_\epsilon\left(\langle\mathbf{W}',\Phi_K(\mathbf{x})\rangle,y(\mathbf{x})\right)\right]\!\!\right] \leq \frac{1}{2C\gamma^2}+\epsilon_0$. Setting $C=\frac{1}{2\epsilon_1\gamma^2}$ gives us our result.

We can use variational techniques to extend this to non-discrete distributions as well.

Part 2: Tightness: The tight example that we provide is an adaptation of the example given for large margin classification in [9]. However, our analysis differs from that of [9], partly necessitated by our choice of loss function.

Consider the following regression problem: $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} \subset \mathbb{R}^3$, $\mathcal{D} = \left\{\frac{1}{2} - \epsilon, \epsilon, \epsilon, \frac{1}{2} - \epsilon\right\}$, $y = \{+1, +1, -1, -1\}$

$$\begin{aligned}
\mathbf{x}_1 &= \left(\gamma, \gamma, \sqrt{1-2\gamma^2}\right) \\
\mathbf{x}_2 &= \left(\gamma, -\gamma, \sqrt{1-2\gamma^2}\right) \\
\mathbf{x}_3 &= \left(-\gamma, \gamma, \sqrt{1-2\gamma^2}\right) \\
\mathbf{x}_4 &= \left(-\gamma, -\gamma, \sqrt{1-2\gamma^2}\right)
\end{aligned}$$

Clearly the vector $\mathbf{w} = (1, 0, 0)$ yields a predictor y' with no ϵ-insensitive loss for $\epsilon = 0$ (i.e. $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![\ell_0(y(\mathbf{x}) - y'(\mathbf{x}))]\!] = 0$) at margin γ. Thus the native inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^3$ is a $(0, \gamma)$-good kernel for this particular regression problem.

Now consider any bounded weighing function on $\mathcal{X}$, $w = \{w_1, w_2, w_3, w_4\}$ and analyze the effectiveness of $\langle\cdot,\cdot\rangle$ as a similarity function. The output $\tilde{y}$ of the resulting predictor on the different points is given by $\tilde{y}_i = \sum_{j=1}^{4} p_j w_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle$.

In particular, consider the output on the *heavy* points $\mathbf{x}_1$ and $\mathbf{x}_4$ (note that the analysis in [9] considers the *light* points $\mathbf{x}_2$ and $\mathbf{x}_3$ instead). We have

$$\begin{aligned}
\tilde{y}_1 &= \left(\frac{1}{2} - \epsilon\right) w_1 + \epsilon\left(1 - 2\gamma^2\right)(w_2 + w_3) + \left(\frac{1}{2} - \epsilon\right) w_4\left(1 - 4\gamma^2\right) = a + \left(\frac{1}{2} - \epsilon\right)(w_1 + b w_4) \\
\tilde{y}_4 &= \left(\frac{1}{2} - \epsilon\right) w_1\left(1 - 4\gamma^2\right) + \epsilon\left(1 - 2\gamma^2\right)(w_2 + w_3) + \left(\frac{1}{2} - \epsilon\right) w_4 = a + \left(\frac{1}{2} - \epsilon\right)(b w_1 + w_4)
\end{aligned}$$

for $a = \epsilon\left(1 - 2\gamma^2\right)(w_2 + w_3), b = \left(1 - 4\gamma^2\right)$. The main idea behind this choice is that the difference in the value of the predictor on these points is only due to the values of w_1 and w_4. Since the true values at these points are very different, this should force w_1 and w_4 to take large values unless a large error is incurred. To formalize this argument we lower bound the expected $\ell_0(\cdot)$ loss of this predictor by the loss incurred on these heavy points.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![\ell_0(y(\mathbf{x}) - \tilde{y}(\mathbf{x}))]\!] &\geq \left(\frac{1}{2} - \epsilon\right)\left(\ell_0(y(\mathbf{x}_1) - \tilde{y}(\mathbf{x}_1)) + \ell_0(y(\mathbf{x}_4) - \tilde{y}(\mathbf{x}_4))\right) \\
&= \left(\frac{1}{2} - \epsilon\right)\left(|1 - \tilde{y}(\mathbf{x}_1)| + |-1 - \tilde{y}(\mathbf{x}_4)|\right) \\
&\geq \left(\frac{1}{2} - \epsilon\right)\left(2 - \tilde{y}(\mathbf{x}_1) + \tilde{y}(\mathbf{x}_4)\right) \\
&= \left(\frac{1}{2} - \epsilon\right)\left(2 - \left(\frac{1}{2} - \epsilon\right)(1 - b)(w_4 - w_1)\right) \\
&= \left(\frac{1}{2} - \epsilon\right)\left(2 - \left(\frac{1}{2} - \epsilon\right)\left(4\gamma^2\right)(w_4 - w_1)\right)
\end{aligned}$$

where in the second step we use the fact that $\ell_0(x) = |x|$ and in the third step we used the fact that $|a| + |b| \geq a - b$. Thus, in order to have expected error at most ϵ_1, we require

$$w_4 - w_1 \geq \frac{1}{4\gamma^2}\left(2 - \frac{\epsilon_1}{\frac{1}{2} - \epsilon}\right)\frac{1}{\frac{1}{2} - \epsilon} = \frac{1}{4\epsilon_1\gamma^2}$$

for the setting $\epsilon = \frac{1}{2} - \epsilon_1$. Thus we have $|w_1| + |w_4| \geq w_4 - w_1 \geq \frac{1}{4\epsilon_1\gamma^2}$ which implies $\max(|w_1|, |w_4|) \geq \frac{1}{8\epsilon_1\gamma^2}$ which proves the result.

Appendix D Sparse Regression with Similarity functions

Our utility proof proceeds in three steps. In the first step we project our learning problem, via the landmarking step given in Step 1 of Algorithm 1, to a linear landmarked space and show that the

Algorithm 2 Sparse regression [10]

Input: A β-smooth loss function $\ell(\cdot,\cdot)$, regularization parameter C_W used in Equation 2, error tolerance ϵ
Output: A sparse predictor $\hat{\mathbf{w}}$ with bounded loss

$k \leftarrow \left\lceil \frac{8C_W^2}{\epsilon^2} \right\rceil$, $\mathbf{w}^{(0)} = \mathbf{0}$
for $t = 1$ **to** k **do**
  $\boldsymbol{\theta}^{(t)} \leftarrow \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}^{(t)}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}} \left[\!\!\left[\frac{\partial}{\partial \mathbf{w}} \ell\left(\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle, y(\mathbf{x}) \right) \right]\!\!\right]$
  $r_t = \arg\max_{j \in d} \left| \boldsymbol{\theta}_j^{(t)} \right|$
  $\delta_t = \left\langle \boldsymbol{\theta}^{(t)}, \mathbf{w}^{(t)} \right\rangle + C_W \left\| \boldsymbol{\theta}^{(t)} \right\|_\infty$
  $\eta_t = \min\left\{1, \frac{\delta_t}{4C_W^2 \beta}\right\}$
  $\mathbf{w}^{(t+1)} \leftarrow (1-\eta_t)\,\mathbf{w}^{(t)} + \eta_t \text{sign}\left(-\boldsymbol{\theta}_{r_t}^{(t)}\right) C_W \mathbf{e}^{r_t}$
  if $\delta_t \leq \epsilon$ **then**
    return $\mathbf{w}^{(t)}$
  end if
end for
return $\mathbf{w}^{(k)}$

landmarked space admits a sparse linear predictor with bounded ϵ-insensitive loss. This is formalized in Theorem 8 which we restate for convenience.

Theorem 21 (Theorem 8 restated)**.** *Given a similarity function that is (ϵ_0, B, τ)-good for a regression problem, there exists a randomized map $\Psi : \mathcal{X} \to \mathbb{R}^d$ for $d = \mathcal{O}\left(\frac{B^2}{\tau\epsilon_1^2} \log \frac{1}{\delta}\right)$ such that with probability at least $1-\delta$, there exists a linear operator $\tilde{f} : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ over $\mathbb{R}^d$ such that $\|\mathbf{w}\|_1 \leq B$ with ϵ-insensitive loss bounded by $\epsilon_0 + \epsilon_1$. Moreover, with the same confidence we have $\|\mathbf{w}\|_0 \leq \frac{3d\tau}{2}$.*

Proof. The proof of this theorem essentially parallels that of [15, Theorem 8] but diverges later since the aim there is to preserve margin violations whereas we wish to preserve loss under the absolute loss function. Sample d *landmark points* $\mathcal{L} = \{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$ from the distribution $\mathcal{D}$ and construct the map $\Psi_{\mathcal{L}} : \mathbf{x} \mapsto (K(\mathbf{x}, \mathbf{x}_1), \ldots, K(\mathbf{x}, \mathbf{x}_d))$ and consider the linear operator $\tilde{f} : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ with $\mathbf{w}_i = \frac{w(\mathbf{x}_i) R(\mathbf{x}_i)}{d_{\text{info}}}$ where $d_{\text{info}} = \sum_{i=1}^{d} R(\mathbf{x}_i)$ is the number of informative landmarks. In the following we will refer to $\tilde{f}$ and $\mathbf{w}$ interchangeably. This ensures that $\left\|\tilde{f}\right\|_1 := \|\mathbf{w}\|_1 \leq B$. Note that we have chosen an L_1 normalized weight vector instead of an L_2 normalized one like we had in Lemma 15. This is due to a subsequent use of sparsity promoting regularizers whose analysis requires the existence of bounded L_1 norm predictors.

Using the arguments given for Lemma 15 and Theorem 5, we can show that if $d_{\text{info}} = \Omega\left(\frac{B^2}{\epsilon_1^2} \log \frac{1}{\delta}\right)$ (i.e. if we have collected enough informative landmarks), then we are done. However, the Chernoff bound (lower tail) tells us that for $d = \Omega\left(\frac{B^2}{\tau\epsilon_1^2} \log \frac{1}{\delta}\right)$, this will happen with probability $1-\delta$. Moreover, the Chernoff bound (upper tail) tells us that, simultaneously we will also have $d_{\text{inf}} \leq \frac{3d\tau}{2}$. Together these prove the claim. □

Note that the number of informative landmarks required is, up to constant factors, the same as the number required in Theorem 5. However, we see that in order to get these many informative landmarks, we have to sample a much larger number number of landmarks. In the following, we shall see how to extract a sparse predictor in the landmarked space with good generalization properties. The following analysis shall assume the the existence of a good predictor on the landmarked space and hence all subsequent results shall be conditioned on the guarantees given by Theorem 8.

D.1 Learning sparse predictors in the landmarked space

We use the *Forward Greedy Selection* algorithm presented in [10] to extract a sparse predictor in the landmarked space. The algorithm is presented in pseudo code form in Algorithm 2. The algorithm can be seen as a (modified) form of orthogonal matching pursuit wherein at each step we add a coordinate to the support of the weight vector. The coordinate is added in a greedy manner so as to provide maximum incremental benefit in terms of lowering the loss. Thus the sparsity of the resulting predictor is bounded by the number of steps for which this algorithm is allowed to run. The algorithm requires that it be used with a *smooth* loss function. A loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is said to be β-smooth if, for all $y, a, b \in \mathbb{R}$, we have

$$\ell(a, y) - \ell(b, y) \leq \frac{\partial}{\partial x}\ell(x, y)\bigg|_{x=b} (a - b) + \frac{\beta(a-b)^2}{2}$$

Unfortunately, this excludes the ϵ-insensitive loss. However it is possible to run the algorithm with a smooth surrogate whose loss can be transferred to ϵ-insensitive loss. Following [10], we choose the following loss function:

$$\tilde{\ell}_\beta(a, b) = \inf_{v \in \mathbb{R}} \left[\frac{\beta}{2} v^2 + \ell_\epsilon (a - v, b) \right]$$

One can, by a mildly tedious case-by-case analysis, arrive at an explicit form for this loss function

$$\tilde{\ell}_\beta(a, b) = \begin{cases} 0 & |a - b| \leq \epsilon \\ \frac{\beta}{2} (|a - b| - \epsilon)^2 & \epsilon < |a - b| < \epsilon + \frac{1}{\beta} \\ |a - b| - \epsilon - \frac{1}{2\beta} & |a - b| \geq \epsilon + \frac{1}{\beta} \end{cases}$$

Note that this loss function is convex as well as differentiable (actually β-smooth) which will be crucial in the following analysis. Moreover, for any a, b we have

$$0 \leq \ell_\epsilon (a, b) - \tilde{\ell}_\beta(a, b) \leq \frac{1}{2\beta} \tag{1}$$

Analysis of Forward Greedy Selection: We need to setup some notation before we can describe the guarantees given for the predictor learned using the Forward Greedy Selection algorithm. Consider a domain $\mathcal{X} \subset \mathbb{R}^d$ for some $d > 0$ and the class of functions $\mathcal{F} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_1 \leq C_W\}$. For any distribution $\mathcal{D}$ on $\mathcal{X}$ and any predictor from $\mathcal{F}$, define $\mathcal{R}_\mathcal{D}(\mathbf{w}) := \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} [\![\ell_\epsilon (\langle \mathbf{w}, \mathbf{x} \rangle, y(\mathbf{x}))]\!]$ and $\tilde{\mathcal{R}}_\mathcal{D}(\mathbf{w}) := \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\tilde{\ell}_\beta(\langle \mathbf{w}, \mathbf{x} \rangle, y(\mathbf{x})) \right]\!\!\right]$. Also let $\bar{\mathbf{w}}$ be the minimizer of the following program

$$\bar{\mathbf{w}} = \mathop{\arg\min}_{\mathbf{w}: \|\mathbf{w}\|_1 \leq C_W} \tilde{\mathcal{R}}_\mathcal{D}(\mathbf{w}) \tag{2}$$

Then [10, Theorem 2.4], when specialized to our case, guarantees that Algorithm 2, when executed with $\tilde{\ell}_\beta(\cdot, \cdot)$ as the loss function for $\beta = \frac{1}{\epsilon_2}$, produces a k-sparse predictor $\hat{\mathbf{x}}$, for $k = \left\lceil \frac{8C_W^2}{\epsilon_2^2} \right\rceil$, with $\|\hat{\mathbf{w}}\|_1 \leq C_W$ such that

$$\tilde{\mathcal{R}}_\mathcal{D}(\hat{\mathbf{w}}) - \tilde{\mathcal{R}}_\mathcal{D}(\bar{\mathbf{w}}) \leq \epsilon_2$$

Thus, if we can show the existence of a good predictor in our space with bounded L_1 norm then this would upper bound the loss incurred by the minimizer of Equation 2 and using [10, Theorem 2.4] we would be done. Note that Theorem 8 does indeed give us such a guarantee which allows us to make the following argument: we are guaranteed of the existence of a predictor $\tilde{f}$ with L_1 norm bounded by B that has ϵ-insensitive loss bounded by $(\epsilon_0 + \epsilon_1)$. Thus if we take $C_W = B$ in Equation 2 and use the left inequality of Equation 1, we get $\tilde{\mathcal{R}}_\mathcal{D}(\bar{\mathbf{w}}) \leq \epsilon_0 + \epsilon_1$. Thus we have $\tilde{\mathcal{R}}_\mathcal{D}(\hat{\mathbf{w}}) \leq \epsilon_0 + \epsilon_1 + \epsilon_2$. Using Equation 1 (right inequality) with $\beta = \frac{1}{\epsilon_2}$, we get $\mathcal{R}_\mathcal{D}(\hat{\mathbf{w}}) \leq \epsilon_0 + \epsilon_1 + 3\epsilon_2/2$.

However it is not possible to give utility guarantees with bounded sample complexities using the above analysis, the reason being that Algorithm 2 requires us to calculate, for any given vector $\mathbf{w}$, the vector $\nabla_\mathbf{w} \tilde{\mathcal{R}}(\mathbf{w}) = \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\frac{\partial}{\partial \mathbf{w}} \tilde{\ell}_\beta(\langle \mathbf{w}, \mathbf{x} \rangle, y(\mathbf{x})) \right]\!\!\right]$ which is infeasible to calculate for a distribution with infinite support since it requires unbounded sample complexities. To remedy we shall, as

suggested by [10], take $\mathcal{D}$ not to be the true distribution over the entire domain $\mathcal{X}$, but rather the *empirical distribution* $\mathcal{D}_{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{x}=\mathbf{x}_i\}}$ for a given sample of training points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Note that the result in [10] holds for any distribution which allows us to proceed as before.

Notice however, that we are yet again faced with the challenge of proving an upper bound on the loss incurred by the minimizer of Equation 2. This we do as follows: the predictor $\tilde{f}$ defined in Theorem 8 has expected ϵ-insensitive loss over the entire domain bounded by $\epsilon_0 + \epsilon_1$. Hence it will, with probability greater than $(1-\delta)$, have at most $\epsilon_0 + \epsilon_1 + \mathcal{O}\left(\frac{B}{\sqrt{n}}\right)$ loss on a random sample of n points by an application of Hoeffding's inequality. Thus we have $\tilde{\mathcal{R}}_{\mathcal{D}_{\text{emp}}}(\bar{\mathbf{w}}) \leq \epsilon_0 + \epsilon_1 + \mathcal{O}\left(\frac{B}{\sqrt{n}}\right)$ with high probability.

The main difference in this analysis shall be that the guarantee on $\hat{\mathbf{w}}$ we get will be on its *training loss* rather than its true loss, i.e. we will have $\mathcal{R}_{\mathcal{D}_{\text{emp}}}(\hat{\mathbf{w}}) \leq \epsilon_0 + \epsilon_1 + \mathcal{O}\left(\frac{B}{\sqrt{n}}\right) + \epsilon_2$. However since Algorithm 2 guarantees $\|\hat{\mathbf{w}}\|_1 \leq C_W = B$, we can still hope to bound its generalization error. More specifically, Lemma 22, given below, shows that with probability greater than $(1-\delta)$ over the choice of training points we will have, for all $\mathbf{w} \in \mathbb{R}^d$, $\mathcal{R}_{\mathcal{D}}(\mathbf{w}) - \mathcal{R}_{\mathcal{D}_{\text{emp}}}(\mathbf{w}) \leq \tilde{\mathcal{O}}\left(\frac{B}{\sqrt{n}}\right)$ where the $\tilde{\mathcal{O}}(\cdot)$ notation hides certain log factors.

Lemma 22 (Risk bounds for sparse linear predictors [13])**.** *Consider a real-valued prediction problem y over a domain $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_\infty \leq C_X\} \subset \mathbb{R}^d$ and a linear learning model $\mathcal{F} : \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\|_0 \leq k, \|\mathbf{w}\|_1 \leq C_W\}$ under some fixed loss function $\ell(\cdot,\cdot)$ that is C_L-Lipschitz in its second argument. For any $f \in \mathcal{F}$, let $\mathcal{L}_f = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![\ell(f(\mathbf{x}), y(\mathbf{x}))]\!]$ and $\hat{\mathcal{L}}_f^n$ be the empirical loss on a set of n i.i.d. chosen points, then we have, with probability greater than $(1-\delta)$,*

$$\sup_{f\in\mathcal{F}}\left(\mathcal{L}_f - \hat{\mathcal{L}}_f^n\right) \leq 2C_L C_X C_W\sqrt{\frac{2\log(2d)}{n}} + C_L C_X C_W\sqrt{\frac{\log(1/\delta)}{2n}}$$

Proof. The result for non-sparse vectors, that applies here as well, follows in a straightforward manner from [13, Theorem 1, Example 3.1(2)] and [23] which we reproduce for completeness. Since the L_1 and L_∞ norms are dual to each other, for any $\mathbf{w} \in (\mathbb{R}^+)^d$ such that $\|\mathbf{w}\|_1 = B$ and any $\mu \in \Delta^d$, where Δ^d is the probability simplex in d dimensions, the Kullback-divergence function $\mathrm{KL}\left(\frac{\mathbf{w}}{B} \middle\| \mu\right)$ is $\frac{1}{B^2}$-strongly convex with respect to the L_1 norm. We can remove the positivity constraints on the coordinates of $\mathbf{w}$ by using the standard method of introducing additional dimensions that encode negative components of the (signed) weight vector.

Using [13, Theorem 1], thus, we can bound the Rademacher complexity of the function class $\mathcal{F}$ as $\mathcal{R}_n(\mathcal{F}) \leq C_X C_W\sqrt{\frac{2\log 2d}{n}}$. Next, using the Lipschitz properties of the loss function, a result from [23] allows us to bound the excess error by $2C_L\mathcal{R}_n(\mathcal{F}) + C_L C_X C_W\sqrt{\frac{\log(1/\delta)}{2n}}$. The result then follows. □

Thus, by applying a union bound, with probability at least $(1-2\delta)$, we will choose a training set such that $\tilde{f}$, and consequently $\bar{\mathbf{w}}$, has bounded loss on that set as well as the uniform convergence guarantee of Lemma 22 will hold. Then we can bound the true loss of the predictor returned by Algorithm 2 as

$$\mathcal{R}_{\mathcal{D}}(\hat{\mathbf{w}}) \leq \mathcal{R}_{\mathcal{D}_{\text{emp}}}(\hat{\mathbf{w}}) + \tilde{\mathcal{O}}\left(\frac{B}{\sqrt{n}}\right) \leq \epsilon_0 + \epsilon_1 + \epsilon_2 + \tilde{\mathcal{O}}\left(\frac{B}{\sqrt{n}}\right)$$

where the first inequality uses the uniform convergence guarantee and the second inequality holds conditional on $\tilde{f}$ having bounded loss on a given training set. The final guarantee is formally given in Theorem 9.

Note that using Lemma 16 here would at best guarantee a decay of $\mathcal{O}\left(\sqrt{\frac{d}{n}}\right)$. Transferring ϵ-insensitive loss to absolute loss requires an addition of ϵ. Using all the results given above, we can now give a proof for Theorem 9 which we restate for convenience.

Theorem 23 (Theorem 9 restated)**.** *Every similarity function that is (ϵ_0, B, τ)-good for a regression problem with respect to the insensitive loss function $\ell_\epsilon(\cdot,\cdot)$ is $(\epsilon_0+\epsilon)$-useful with respect to absolute loss as well; with the dimensionality of the landmarked space being bounded by $\mathcal{O}\left(\frac{B^2}{\tau\epsilon_1^2}\log\frac{1}{\delta}\right)$ and the labeled sampled complexity being bounded by $\mathcal{O}\left(\frac{B^2}{\epsilon_1^2}\log\frac{B}{\epsilon_1\delta}\right)$. Moreover, this utility can be achieved by an $\mathcal{O}(\tau)$-sparse predictor on the landmarked space.*

Proof. Using Theorem 8, we first bound the excess loss due to landmarking by $32B\sqrt{\frac{\log(1/\tau\delta)}{d}}$. Next we set up the (dummy) Ivanov regularized regression problem (given in Equation 2) with the training loss being the objective and regularization parameter $C_W = B$. The training loss incurred by the minimizer of that problem $\mathbf{w}_{\text{inter}}$ is, with probability at least $(1-\delta)$, bounded by $\hat{\mathcal{L}}(\mathbf{w}_{\text{inter}}) \leq \epsilon_0 + 32B\sqrt{\frac{\log(1/\delta)}{\tau d}} + B\sqrt{\frac{\log(1/\delta)}{n}}$ due to the guarantees of Theorem 8. Next, we run the Forward Greedy Selection algorithm of [10] (specialized to our case in Algorithm 2) and obtain another predictor $\hat{\mathbf{w}}$ with L_1 norm bounded by B that has empirical error at most $\hat{\mathcal{L}}(\hat{\mathbf{w}}) \leq \hat{\mathcal{L}}(\mathbf{w}_{\text{inter}}) + \sqrt{\frac{18B^2}{k}}$. Finally, using Lemma 22, we bound the true ϵ-insensitive loss incurred by $\hat{\mathbf{w}}$ by $\hat{\mathcal{L}}(\hat{\mathbf{w}}) + 2B\sqrt{\frac{2\log(2d)}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}$. Adding ϵ to convert this loss to absolute loss we get that with probability at most $(1-3\delta)$, we will output a k-sparse predictor in a d-dimensional space with absolute regression loss at most

$$\epsilon_0 + 32B\sqrt{\frac{\log(1/\delta)}{\tau d}} + \sqrt{\frac{18B^2}{k}} + 2B\sqrt{\frac{2\log(2d)}{n}} + 2B\sqrt{\frac{\log(1/\delta)}{2n}} + \epsilon \qquad \square$$

We note that Forward Greedy Selection gives $\mathcal{O}\left(\frac{1}{k}\right)$ error rates, which are much better, if the loss function being used is smooth. This can be achieved by using squared loss $\ell_{\text{sq}}(a,b) = (a-b)^2$ as the surrogate. However we note that assuming goodness of the similarity function in terms of squared loss would impose strictly stronger conditions on the learning problem. This is because $\mathbb{E}[\![\ell_{\text{sq}}(a,b)]\!] = \sup(a-b)\cdot\mathbb{E}[\![|a-b|]\!]$ and thus, under boundedness conditions, squared loss is bounded by a constant times the absolute loss but it is not possible to bound absolute loss (or ϵ-insensitive loss) as a constant multiple of the squared loss since there exist distributions such that $\mathbb{E}[\![|a-b|]\!] = \Omega\left(\frac{1}{\inf(|a-b|)}\cdot\mathbb{E}[\![\ell_{\text{sq}}(a,b)]\!]\right)$ and $\frac{1}{\inf(|a-b|)}$ can diverge.

Below we prove admissibility results for the sparse learning model.

D.2 Proof of Theorem 10

To prove the first part, construct a new weight function $\tilde{w}(\mathbf{x}) = \text{sign}(w(\mathbf{x}))\cdot\bar{w}$. Note that we have $|\tilde{w}(\mathbf{x})| \leq \bar{w} \leq B$. Also construct the choice function as follows: for any $\mathbf{x}$, let $\mathbb{P}[R(\mathbf{x})=1|\mathbf{x}] = \frac{|w(\mathbf{x})|}{B}$. This gives us $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![R(\mathbf{x})]\!] = \frac{\bar{w}}{B}$. Then for any $\mathbf{x}$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[\![\tilde{w}(\mathbf{x}')K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}')]\!] &= \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}\left[\!\!\left[\text{sign}(w(\mathbf{x}))\,\bar{w}K(\mathbf{x},\mathbf{x}')\frac{|w(\mathbf{x})|}{B}\right]\!\!\right] \Big/ \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[R(\mathbf{x})=1] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}\left[\!\!\left[w(\mathbf{x})K(\mathbf{x},\mathbf{x}')\frac{\bar{w}}{B}\right]\!\!\right] \Big/ \left(\frac{\bar{w}}{B}\right) \\
&= \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[\![w(\mathbf{x}')K(\mathbf{x},\mathbf{x}')]\!]
\end{aligned}$$

Since $f(\mathbf{x}) = \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[\![w(\mathbf{x}')K(\mathbf{x},\mathbf{x}')]\!]$ has small ϵ-insensitive loss by (ϵ_0, B)-goodness of K, we have our result. To prove the second part, construct a new weight function $\tilde{w}(\mathbf{x}) = \frac{w(\mathbf{x})}{\tau}\mathbb{P}[R(\mathbf{x})=1|\mathbf{x}]$.

Note that we have $|\tilde{w}(\mathbf{x})| \leq \frac{B}{\tau}$. Then for any $\mathbf{x}$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[\![\tilde{w}(\mathbf{x}')K(\mathbf{x},\mathbf{x}')]\!] &= \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}\left[\!\!\left[\frac{w(\mathbf{x}')}{\tau}R(\mathbf{x}')K(\mathbf{x},\mathbf{x}')\right]\!\!\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}\left[\!\!\left[\frac{w(\mathbf{x}')}{\tau}K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}')\right]\!\!\right]\mathop{\mathbb{P}}_{\mathbf{x}'\sim\mathcal{D}}[R(\mathbf{x}')=1] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[\![w(\mathbf{x}')K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}')]\!]
\end{aligned}$$

Since $f(\mathbf{x}) = \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[\![w(\mathbf{x}')K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}')]\!]$ has small ϵ-insensitive loss by (ϵ_0, B, τ)-goodness of K, we have our result.

Using the above result we get out admissibility guarantee.

Corollary 24. *Every PSD kernel that is (ϵ_0, γ)-good for a regression problem is, for any $\epsilon_1 > 0$, $\left(\epsilon_0 + \epsilon_1, \mathcal{O}\left(\frac{1}{\epsilon_1\gamma^2}\right), 1\right)$-good as a similarity function as well.*

The above result is rather weak with respect to the sparsity parameter τ since we have made no assumptions on the distribution of the dual variables α_i, α_i^* in the proof of Theorem 6 which is why we are forced to use the (weak) inequality $\frac{\bar{w}}{B} \leq 1$. Any stronger assumptions on the kernel goodness shall also strengthen this admissibility result.

Appendix E Ordinal Regression

In this section we give missing utility and admissibility proofs for the similarity-based learning model for ordinal regression. But before we present the analysis of our model, we give below, an analysis of algorithms that choose to directly reduce the ordinal regression problem to real-valued regression. The analysis will serve as motivation that will help us define our goodness criteria.

E.1 Reductions to real valued regression

One of the simplest learning algorithms for the problem of ordinal regression involves a reduction to real-valued regression [17, 16] where we modify our goal to that of learning a real valued function f which we then threshold using a set of thresholds $\{b_i\}_{i=1}^r$ with $b_1 = -\infty$ to get discrete labels as shown below

$$y_f(\mathbf{x}) = \mathop{\arg\max}_{i\in[r]}\{b_i : f(\mathbf{x}) \geq b_i\}$$

These thresholds may themselves be learned or fixed apriori. A simple choice for these thresholds is $b_i = i - 1$ for $i > 1$. It is easy to show (using a result in [17]) that for the fixed thresholds specified above, we have for all $f : \mathcal{X} \to \mathbb{R}$,

$$\begin{aligned}
\ell_{\text{ord}}\left(y_f(\mathbf{x}), y(\mathbf{x})\right) &\leq \min\left\{2\left|f(\mathbf{x}) - y(\mathbf{x})\right|, \left|f(\mathbf{x}) - y(\mathbf{x})\right| + \frac{1}{2}\right\} \\
&\leq \min\left\{2\ell_\epsilon\left(f(\mathbf{x}) - y(\mathbf{x})\right) + 2\epsilon, \ell_\epsilon\left(f(\mathbf{x}) - y(\mathbf{x})\right) + \epsilon + \frac{1}{2}\right\}
\end{aligned}$$

where in the last step we use the fact that $|x| - \epsilon \leq \ell_\epsilon(x) \leq |x|$.

It is tempting to use this reduction along with guarantees given for real-valued regression to directly give generalization bounds for ordinal regression. To pursue this further, we need a notion of a good similarity function which we give below:

Definition 25. *A similarity function K is said to be (ϵ_0, B)-good for an ordinal regression problem $y : \mathcal{X} \to [r]$ if for some bounded weight function $w : \mathcal{X} \to [-B, B]$, the following predictor, when subjected to fixed thresholds, has expected ordinal regression error at most ϵ_0*

$$f : \mathbf{x} \mapsto \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[\![w(\mathbf{x}')K(\mathbf{x},\mathbf{x}')]\!]$$

i.e. $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![|y_f(\mathbf{x}) - y(\mathbf{x})|]\!] < \epsilon_0$.

From the definition of the thresholding scheme used to define y_f from f, it is clear that $|f(\mathbf{x}) - y(\mathbf{x})| \leq |y_f(\mathbf{x}) - y(\mathbf{x})| + \frac{1}{2}$. Since we have $\ell_\epsilon(x) \leq |x|$ for any $\epsilon \geq 0$, we have $\ell_\epsilon(f(\mathbf{x}) - y(\mathbf{x})) \leq |y(\mathbf{x}) - y_f(\mathbf{x})| + \frac{1}{2}$ and thus we have $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\![\ell_\epsilon(f(\mathbf{x}), y(\mathbf{x}))]\!] < \epsilon_0 + \frac{1}{2}$.

Thus, starting with goodness guarantee of the similarity function with respect to ordinal regression, we obtain a guarantee of the goodness of the similarity function K with respect to real-valued regression that satisfies the requirements of Theorem 5. Thus we have the existence of a linear predictor over a low dimensional space with ϵ-insensitive error at most $\epsilon_0 + \frac{1}{2} + \epsilon_1$. We can now argue (using results from [17]) that this real-valued predictor, when subjected to the fixed thresholds, would yield a predictor with ordinal regression error at most

$$\min\left\{2\left(\epsilon_0 + \frac{1}{2} + \epsilon_1\right) + 2\epsilon, \left(\epsilon_0 + \frac{1}{2} + \epsilon_1\right) + \epsilon + \frac{1}{2}\right\} = 1 + \epsilon_0 + \epsilon_1 + \epsilon.$$

However, this is rather disappointing since this implies that the resulting predictor would, on an average, give out labels that are at least one step away from the true label. This forms the intuition behind introducing (soft) margins in the goodness formulation that gives us Definition 12. Below we give proofs for utility and admissibility guarantees for our model for similarity-based ordinal regression.

E.2 Proof of Theorem 13

We use Lemma 15 to construct a landmarked space with a linear predictor $\tilde{f} : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ such that $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\left|\tilde{f}(\Psi(\mathbf{x})) - f(\mathbf{x})\right|\right]\!\!\right] \leq 2\epsilon_1$. As before, we have $\|\mathbf{w}\|_2 \leq B$ and $\sup_{\mathbf{x}\in\mathcal{X}}\{\|\Psi(\mathbf{x})\|\} \leq 1$. In the following, we shall first show bounds on the mislabeling error i.e $\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{y}(\mathbf{x}) \neq y(\mathbf{x})]$. Next, we shall convert these bounds into ordinal regression loss by introducing a *spacing* parameter into the model.

Since the γ-margin loss function is 1-Lipschitz, we get

$$\begin{aligned}\left[\tilde{f}(\Psi(\mathbf{x})) - b_{y(\mathbf{x})}\right]_\gamma &\leq \left[f(\mathbf{x}) - b_{y(\mathbf{x})}\right]_\gamma + 2\epsilon_1 \\ \left[b_{y(\mathbf{x})+1} - \tilde{f}(\Psi(\mathbf{x}))\right]_\gamma &\leq \left[b_{y(\mathbf{x})+1} - f(\mathbf{x})\right]_\gamma + 2\epsilon_1\end{aligned}$$

Which gives us, upon taking expectations on both sides,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\left[\tilde{f}(\Psi(\mathbf{x})) - b_{y(\mathbf{x})}\right]_\gamma + \left[b_{y(\mathbf{x})+1} - \tilde{f}(\Psi(\mathbf{x}))\right]_\gamma\right]\!\!\right] \leq \epsilon_0 + 4\epsilon_1$$

Lemma 15 guarantees the excess loss due to landmarking to be at most $64B\sqrt{\frac{\log(1/\delta)}{d}}$. Moreover, since the γ-margin loss is 1-Lipschitz, Lemma 16 allows us to bound excess loss due to training by $3B\sqrt{\frac{\log(1/\delta)}{n}}$ so that the learned predictor has γ-margin loss at most $\epsilon_0 + \epsilon_1$ for any ϵ_1 given large enough d and n. Now, from the definition of the γ-margin loss it is clear that if the loss is greater than γ then it indicates a mislabeling. Hence, the mislabeling error is bounded by $\frac{\epsilon_0+\epsilon_1}{\gamma}$.

This may be unsatisfactory if $\gamma \ll 1$ - to remedy such situations we show that we can bound the 1-margin loss directly. Starting from $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\left|\tilde{f}(\Psi(\mathbf{x})) - f(\mathbf{x})\right|\right]\!\!\right] < 2\epsilon_1$, we can also deduce

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\!\left[\left[1 - \tilde{f}(\Psi(\mathbf{x})) + b_{y(\mathbf{x})}\right]_+ + \left[1 - b_{y(\mathbf{x})+1} + \tilde{f}(\Psi(\mathbf{x}))\right]_+\right]\!\!\right] \leq \epsilon_0 + 4\epsilon_1$$

We can bound the excess training error for this loss function as well. Since the 1-margin loss directly bounds the mislabeling error, combining the two arguments we get the second part of the claim.

However, the margin losses themselves do not present any bound on the ordinal regression error. This is because, if the thresholds are closely spaced together, then even an instance of gross ordinal regression loss could correspond to very small margin loss. To remedy this, we introduce a *spacing* parameter into the model. We say that a set of thresholds is Δ-spaced if $\min_{i\in[r]}\{|b_i - b_{i+1}|\} \geq \Delta$.

Such a condition can easily be incorporated into the model of [17] as a constraint in the optimization formulation.

Suppose that a given instance has ordinal regression error $\ell_{\text{ord}}\left(\hat{y}(\mathbf{x}), y(\mathbf{x})\right) = k$. This can happen if the point was given a label k labels below (or above) its correct label. Also suppose that the γ-margin error in this case is $\left[\hat{y}(\mathbf{x}) - y(\mathbf{x})\right]_\gamma = h$. Without loss of generality, assume that the point $\mathbf{x}$ of label $k+1$ was given the label 1 giving an ordinal regression loss of $l_{\text{ord}} = k$ (a similar analysis would hold if the point of label 1 were to be given a label $k+1$ by symmetry of the margin loss formulation with respect to left and right thresholds). In this case the value of the underlying regression function must lie between b_1 and b_2 and thus, the margin loss h satisfies $h \geq b_{k+1} + \gamma - b_2 = \gamma + \sum_{i=2}^{k} \left(b_{i+1} - b_i\right) \geq \gamma + (k-1)\,\Delta$. Thus, if the margin loss is at most h, the ordinal regression error must satisfy $\ell_{\text{ord}}\left(\hat{y}(\mathbf{x}), y(\mathbf{x})\right) \leq \frac{\left[\hat{y}(\mathbf{x}) - b_{y(\mathbf{x})}\right]_\gamma + \left[b_{y(\mathbf{x})+1} - \hat{y}(\mathbf{x})\right]_\gamma - \gamma}{\Delta} + 1$. Let $\psi_\Delta(x) = \frac{x+\Delta-1}{\Delta}$. Using the bounds on the γ-margin and 1-margin losses given above, we get the first part of the claim.

In particular, a constraint of $\Delta = 1$ put into an optimization framework ensures that the bounds on mislabeling loss and ordinal regression loss match since $\psi_1(x) = x$ for all x. In general, the cases where the above framework yields a non-trivial bound for the mislabeling error rate, i.e. $\ell_{01} < 1$ (which can always be ensured if $\epsilon_0 < 1$ by taking large enough d and n), also correspond to those where the ordinal regression error rate is also bounded above by 1 since $\sup_{x \in [0,1], \Delta > 0} \left(\psi_\Delta\left(x\right)\right) = 1$.

E.3 Admissibility Guarantees

We begin by giving the kernel goodness criterion which we adapt from existing literature on large margin approaches to ordinal regression. More specifically we use the framework described in [16] for which generalization guarantees are given in [17].

Definition 26. *Call a PSD kernel K (ϵ_0, γ)-good for an ordinal regression problem $y : \mathcal{X} \to [r]$ if there exists $\mathbf{W}^* \in \mathcal{H}_K$, $\|\mathbf{W}^*\| = 1$ and a fixed set of thresholds $\{b_i\}_{i=1}^r$ such that*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\left[b_{y(\mathbf{x})} + 1 - \frac{\langle \mathbf{W}^*, \Phi_K(\mathbf{x}) \rangle}{\gamma} \right]_+ + \left[\frac{\langle \mathbf{W}^*, \Phi_K(\mathbf{x}) \rangle}{\gamma} - b_{y(\mathbf{x})+1} + 1 \right]_+ \right]\!\!\right] < \epsilon_0$$

The above definition exactly corresponds to the EXC formulation put forward by [17] except for the fact that during actual optimization, a strict ordering on the thresholds is imposed explicitly. [17] present yet another model called IMC which does not impose any explicit orderings, rather the ordering emerges out of the minimization process itself. Our model can be easily extended to the IMC formulation as well.

Theorem 27 (Theorem 14 restated)**.** *Every PSD kernel that is (ϵ_0, γ)-good for an ordinal regression problem is also $\left(\gamma_1 \epsilon_0 + \epsilon_1, \mathcal{O}\left(\frac{\gamma_1^2}{\epsilon_1 \gamma^2}\right)\right)$-good as a similarity function with respect to the γ_1-margin loss for any $\gamma_1, \epsilon_1 > 0$. Moreover, for any $\epsilon_1 < \gamma_1/2$, there exists an ordinal regression instance and a corresponding kernel that is $(0, \gamma)$-good for the ordinal regression problem but only (ϵ_1, B)-good as a similarity function with respect to the γ_1-margin loss function for $B = \Omega\left(\frac{\gamma_1^2}{\epsilon_1 \gamma^2}\right)$.*

Proof. We prove the two parts of the result separately.

Part 1: Admissibility: As before, using Lemma 17 it is possible to obtain a vector $\mathbf{W}' = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \Phi_K(\mathbf{x}_i) \in \mathcal{H}_K$ such that $0 \leq \alpha_i, \alpha_i^* \leq p_i C$ (by applying the KKT conditions) and the following holds:

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\left[b_{y(\mathbf{x})} + 1 - \langle \mathbf{W}', \Phi_K(\mathbf{x}) \rangle \right]_+ + \left[\langle \mathbf{W}', \Phi_K(\mathbf{x}) \rangle - b_{y(\mathbf{x})+1} + 1 \right]_+ \right]\!\!\right] < \frac{1}{2C\gamma^2} + \epsilon_0 \tag{3}$$

This allows us to construct a weight function $w_i = \frac{\alpha_i - \alpha_i^*}{p_i}$ such that $|w_i| \leq 2C$ (since we do not have any guarantee that $\alpha_i \alpha_i^* = 0$) and $\mathop{\mathbb{E}}_{\mathbf{x}' \sim \mathcal{D}} \left[\!\!\left[w(\mathbf{x}') K(\mathbf{x}, \mathbf{x}') \right]\!\!\right] = \langle \mathbf{W}', \Phi_K(\mathbf{x}) \rangle$ for all $\mathbf{x} \in \mathcal{X}$.

Denoting $f(\mathbf{x}) := \mathop{\mathbb{E}}_{\mathbf{x}' \sim \mathcal{D}} [\![w(\mathbf{x}') K(\mathbf{x}, \mathbf{x}')]\!]$ for convenience gives us

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\left[f(\mathbf{x}) - b_{y(\mathbf{x})} \right]_1 + \left[b_{y(\mathbf{x})+1} - f(\mathbf{x}) \right]_1 \right]\!\!\right] &= \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\left[1 - f(\mathbf{x}) + b_{y(\mathbf{x})} \right]_+ + \left[1 - b_{y(\mathbf{x})+1} + f(\mathbf{x}) \right]_+ \right]\!\!\right] \\ &\leq \frac{1}{2C\gamma^2} + \epsilon_0 \end{aligned}$$

where in the first step we used $[x]_1 = [1 - x]_+$. Now use the fact $[x]_1 = \frac{1}{\gamma} [\gamma x]_\gamma$ to get the following:

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\left[\gamma_1 f(\mathbf{x}) - \gamma_1 b_{y(\mathbf{x})} \right]_{\gamma_1} + \left[\gamma_1 b_{y(\mathbf{x})+1} - \gamma_1 f(\mathbf{x}) \right]_{\gamma_1} \right]\!\!\right] \leq \frac{\gamma_1}{2C\gamma^2} + \gamma_1 \epsilon_0$$

Note that it is not possible to perform the analysis on the loss function $[\cdot]_\gamma$ directly since using it requires us to scale the threshold values by a factor of γ_1 that makes the result in Equation 3 unusable. Hence we first perform the analysis for $[\cdot]_1$, utilize Equation 3 and then interpret the resulting inequality in terms of $[\cdot]_{\gamma_1}$.

Setting $2C = \frac{\gamma_1}{\epsilon_1 \gamma^2}$, using $w'(\mathbf{x}) = \gamma_1 w(\mathbf{x})$ as weights, using $b'_j = \gamma_1 b_j$ as the thresholds and noting that the new bound on the weights is $|w'_i| \leq 2C\gamma_1$ gives us the result. As before, using variational optimization techniques, this result can be extended to non-discrete distributions as well. □

In particular, setting $\gamma_1 = \gamma$ gives us that any PSD kernel that is (ϵ_0, γ)-good for an ordinal regression problem is also $\left(\gamma\epsilon_0 + \epsilon_1, \frac{1}{\epsilon_1} \right)$-good as a similarity function with respect to the γ-margin loss.

Part 2: Tightness: We adapt our running example (used for proving the lower bound for real regression) for the case of ordinal regression as well. Consider the points with value -1 as having label 1 and those having value $+1$ as having label 2. Clearly, $w = (1, 0, 0)$ along with the thresholds $b_1 = -\infty$ and $b_2 = 0$ establishes the native inner product as a $(0, \gamma)$-good PSD kernel.

Now consider the heavy points yet again and some weight function and threshold b_2 (b_1 is always fixed at $-\infty$) that is supposed to demonstrate the goodness of the inner product kernel as a similarity function. Clearly we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\left[f(\mathbf{x}) - b_{y(\mathbf{x})} \right]_{\gamma_1} + \left[b_{y(\mathbf{x})+1} - f(\mathbf{x}) \right]_{\gamma_1} \right]\!\!\right] &\geq \left(\frac{1}{2} - \epsilon \right) \left(\left[f(\mathbf{x}_1) - b_2 \right]_{\gamma_1} + \left[b_2 - f(\mathbf{x}_4) \right]_{\gamma_1} \right) \\ &= \left(\frac{1}{2} - \epsilon \right) \left(\left[\gamma_1 - f(\mathbf{x}_1) + b_2 \right]_+ + \left[\gamma_1 - b_2 + f(\mathbf{x}_4) \right]_+ \right) \\ &\geq \left(\frac{1}{2} - \epsilon \right) \left(2\gamma_1 - f(\mathbf{x}_1) + f(\mathbf{x}_4) \right) \\ &= \left(\frac{1}{2} - \epsilon \right) \left(2\gamma_1 - \left(\frac{1}{2} - \epsilon \right) (1 - b) (w_4 - w_1) \right) \\ &= \left(\frac{1}{2} - \epsilon \right) \left(2\gamma_1 - \left(\frac{1}{2} - \epsilon \right) \left(4\gamma^2 \right) (w_4 - w_1) \right) \end{aligned}$$

where in the third step we have used the fact that $[a]_+ + [b]_+ \geq a + b$. Thus, in order to have expected error at most ϵ_1, we must have

$$w_4 - w_1 \geq \frac{1}{4\gamma^2} \left(2\gamma_1 - \frac{\epsilon_1}{\frac{1}{2} - \epsilon} \right) \frac{1}{\frac{1}{2} - \epsilon} = \frac{\gamma_1^2}{4\epsilon_1 \gamma^2}$$

by setting $\epsilon = \frac{1}{2} - \frac{\epsilon_1}{\gamma_1}$ which then proves the result after applying an averaging argument.

Appendix F Ranking

The problem of ranking stems from the need to sort a set of items based on their relevance. In the model considered here, each ranking instance is composed of m documents (pages) $(p_1, \ldots, p_m)$

from some universe $\mathcal{P}$ along with their relevance to some particular query $q \in \mathcal{Q}$ that are given as relevance scores from some set $\mathcal{R} \subset \mathbb{R}$. Thus we have $\mathcal{X} = \mathcal{Q} \times \mathcal{P}^m$ with each instance $\mathbf{x} \in \mathcal{X}$ being provided with a relevance vector $r(\mathbf{x}) = \mathcal{R}^m$. Let the i^{th} query-document pair of a ranking instance $\mathbf{x}$ be denoted by $\mathbf{z}_i \in \mathcal{Q} \times \mathcal{P}$. For any $\mathbf{z} = (p, q) \in \mathcal{P} \times \mathcal{Q}$, let $r(\mathbf{z}) \in \mathbb{R}$ denote the true relevance of document p to query q.

For any relevance vector $\mathbf{r} \in \mathcal{R}^m$, let $\bar{\mathbf{r}}$ be the vector with elements of $\mathbf{r}$ sorted in descending order and $\pi_{\mathbf{r}}$ be the permutation that this sorting induces. For any permutation π, $\pi(i)$ shall denote the index given to the index i under π. Although the desired output of a ranking problem is a permutation, we shall follow the standard simplification [27] of requiring the output to be yet another relevance vector $\mathbf{s}$ with the permutation $\pi_{\mathbf{s}}$ being considered as the actual output. This converts the ranking problem into a vector-valued regression problem.

We will take the true loss function $\ell_{\text{actual}}(\cdot, \cdot)$ to be the popular NDCG loss function [28] defined below

$$\ell_{\text{NDCG}}(\mathbf{s}, \mathbf{r}) = -\frac{1}{\|G(\mathbf{r})\|_D} \sum_{i=1}^{m} \frac{G(\mathbf{r}(i))}{F(\pi_{\mathbf{s}}(i))}$$

where $\|\mathbf{r}\|_D = \max_{\pi \in S_m} \sum_{i=1}^{m} \frac{\mathbf{r}(i)}{F(\pi(i))}$, $G(r) = 2^r - 1$ is the growth function and $F(t) = \log(1 + t)$ is the decay function.

For the surrogate loss functions ℓ_K and ℓ_S, we shall use the squared loss function $\ell_{\text{sq}}(\mathbf{s}, \mathbf{r}) = \|\mathbf{s} - \mathbf{r}\|_2^2$. We shall overload notation to use $\ell_{\text{sq}}(\cdot, \cdot)$ upon reals as well. For any vector $\mathbf{r} \in \mathcal{R}^m$, let $\eta(\mathbf{r}) := \frac{G(\mathbf{r})}{\|G(\mathbf{r})\|_D}$ and let $\mathbf{r}_i$ denote its i^{th} coordinate.

Due to the decomposable nature of the surrogate loss function, we shall require kernels and similarity functions to act over query-document pairs i.e. $K : (\mathcal{P} \times \mathcal{Q}) \times (\mathcal{P} \times \mathcal{Q}) \to \mathbb{R}$. This also coincides with a common feature extraction methodology (see for example [27, 29]) where every query-document pair is processed to yield a feature vector. Consequently, all our goodness definitions shall loosely correspond to the ability of a kernel/similarity to accurately predict the true relevance scores for a given query-document pair. We shall assume ranking instances to be generated by the sampling of a query $q \sim \mathcal{D}_{\mathcal{Q}}$ followed by m independent samples of documents from the (conditional) distribution $\mathcal{D}_{\mathcal{P}|q}$. The distribution over ranking instances is then a product distribution $\mathcal{D} = \mathcal{D}_{\mathcal{X}} = \mathcal{D}_{\mathcal{Q}} \times \underbrace{\mathcal{D}_{\mathcal{P}|q} \times \mathcal{D}_{\mathcal{P}|q} \times \ldots \times \mathcal{D}_{\mathcal{P}|q}}_{m \text{ times}}$. A key consequence of this generative mechanism is that the i^{th} query-document pair of a random ranking instance, for any fixed i, is a random query-document instance selected from the distribution $\mu := \mathcal{D}_{\mathcal{Q}} \times \mathcal{D}_{\mathcal{P}|q}$.

Definition 28. *A similarity function K is said to be (ϵ_0, B)-good for a ranking problem $y : \mathcal{X} \to S_m$ if for some bounded weight function $w : \mathcal{P} \times \mathcal{Q} \to [-B, B]$, for any ranking instance $\mathbf{x} = (q, p_1, p_2, \ldots, p_m)$, if we define $f : \mathcal{X} \to \mathbb{R}^m$ as*

$$f_i := \mathop{\mathbb{E}}_{\mathbf{z} \sim \mu} [\![w(\mathbf{z}) K(\mathbf{z}_i, \mathbf{z})]\!]$$

where $\mathbf{z}_i = (p_i, q)$, then we have $\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} [\![\ell_{sq}(f(\mathbf{x}), \eta(r(\mathbf{z})))]\!] < \epsilon_0$.

Definition 29. *A PSD kernel K is said to be (ϵ_0, γ)-good for a ranking problem $y : \mathcal{X} \to S_m$ if there exists $\mathbf{W}^* \in \mathcal{H}_K$, $\|\mathbf{W}^*\| = 1$ such that if for any ranking instance $\mathbf{x} = (q, p_1, p_2, \ldots, p_m)$, if, for any $\mathbf{W} \in \mathcal{H}_K$, when we define $f(\,\cdot\,; \mathbf{W}) : \mathcal{X} \to \mathbb{R}^m$ as*

$$f_i(\mathbf{x}; \mathbf{W}) = \frac{\langle \mathbf{W}, \Phi_K(\mathbf{z}_i) \rangle}{\gamma}$$

where f_i is the i^{th} coordinate of the output of f and $\mathbf{z}_i = (p_i, q)$, then we have $\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} [\![\ell_{sq}(f(\mathbf{x}; \mathbf{W}^*), \eta(r(\mathbf{z})))]\!] < \epsilon_0$.

The choice of this surrogate is motivated by consistency considerations. We would ideally like a minimizer of the surrogate loss to have bounded actual loss as well. Using results from [27], it can be shown that the above defined surrogate is not only consistent, but that excess loss in terms of

this surrogate can be transferred to excess loss in terms of $\ell_{\text{NDCG}}(\cdot,\cdot)$, a very desirable property. Although [27] shows this to be true for a whole family of surrogates, we chose $\ell_{\text{sq}}(\cdot,\cdot)$ for its simplicity. All our utility arguments carry forward to other surrogates defined in [27] with minimal changes.

We move on to prove utility guarantees for the given similarity learning model.

Theorem 30. *Every similarity function that is (ϵ_0, B)-good for a ranking problem for m-documents with respect to squared loss is $\mathcal{O}\left(\sqrt{\frac{m}{\log m}} \cdot \sqrt{\epsilon_0}\right)$-useful with respect to NDCG loss.*

Proof. As before, we use Lemma 15 to construct a landmarked space with a linear predictor $\tilde{f} : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ such that $\mathop{\mathbb{E}}_{\mathbf{z}\sim\mu}\left[\!\left[\left|\tilde{f}\left(\Psi\left(\mathbf{z}\right)\right) - f(\mathbf{z})\right|\right]\!\right] \leq 2\epsilon_1$. We have $\|\mathbf{w}\|_2 \leq B$ and $\sup_{\mathbf{x}\in\mathcal{X}}\{\|\Psi(\mathbf{x})\|\} \leq 1$. Now lets overload notation to denote by $\Psi(\mathbf{x})$ the concatenation of the images of the m document-query pairs in $\mathbf{x}$ under $\Psi(\cdot)$ and by $\tilde{f}(\Psi(\mathbf{x}))$, the m-dimensional vector obtained by applying $\tilde{f}$ to each of the m components of $\Psi(\mathbf{x})$.

Since the squared loss function is $2B$-Lipschitz in its first argument in the region of interest, we get

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_{\text{sq}}\left(\tilde{f}(\Psi(\mathbf{x})), \eta(r(\mathbf{x}))\right)\right]\!\right] &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\sum_{i=1}^{m} \ell_{\text{sq}}\left(\tilde{f}(\Psi(\mathbf{z}_i)), \eta(r(\mathbf{x}))_i\right)\right]\!\right] \\
&= \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_{\text{sq}}\left(\tilde{f}(\Psi(\mathbf{z}_i)), \eta(r(\mathbf{x}))_i\right)\right]\!\right] \\
&= \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_{\text{sq}}\left(f(\mathbf{z}_i), \eta(r(\mathbf{x}))_i\right)\right]\!\right] + \\
&\quad \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_{\text{sq}}\left(\tilde{f}(\Psi(\mathbf{z}_i)), \eta(r(\mathbf{x}))_i\right) - \ell_{\text{sq}}\left(f(\mathbf{z}_i), \eta(r(\mathbf{x}))_i\right)\right]\!\right] \\
&\leq \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_{\text{sq}}\left(f(\mathbf{z}_i), \eta(r(\mathbf{x}))_i\right)\right]\!\right] + 2B\sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\left|\tilde{f}(\Psi(\mathbf{z}_i)) - f(\mathbf{z}_i)\right|\right]\!\right] \\
&= \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_{\text{sq}}\left(f(\mathbf{z}_i), \eta(r(\mathbf{x}))_i\right)\right]\!\right] + 2B\sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{z}\sim\mu}\left[\!\left[\left|\tilde{f}(\Psi(\mathbf{z})) - f(\mathbf{z})\right|\right]\!\right] \\
&\leq \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_{\text{sq}}\left(f(\mathbf{z}_i), \eta(r(\mathbf{x}))_i\right)\right]\!\right] + 4Bm\epsilon_1 \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\sum_{i=1}^{m} \ell_{\text{sq}}\left(f(\mathbf{z}_i), \eta(r(\mathbf{x}))_i\right)\right]\!\right] + 4Bm\epsilon_1 \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\!\left[\ell_{\text{sq}}\left(f(\mathbf{x}), \eta(r(\mathbf{x}))\right)\right]\!\right] + 4Bm\epsilon_1 \\
&\leq \epsilon_0 + 4Bm\epsilon_1
\end{aligned}
$$

where $\mathbf{x} = (q, p_1, \ldots, p_m)$ and $\mathbf{z}_i = (p_i, q)$. In the first and the last but one step we have used decomposability of the squared loss, in the fourth step we have used Lipschitz properties of the squared loss, in the fifth step we have used properties of the generative mechanism assumed for ranking instances, in the sixth step we have used the guarantee given by Lemma 15. Throughout we have repeatedly used linearity of expectation. This bounds the excess error due to landmarking to d dimensions by $64B^2m^2\sqrt{\frac{\log(1/\delta)}{d}}$ using Lemma 15. Similarly, Lemma 16 also allows us to bound the excess error due to training by $3B^2\sqrt{\frac{\log(1/\delta)}{n}}$ which puts our total squared loss at $\epsilon_0 + \epsilon_1$ for large enough d and n.

We now invoke [27, Theorem 10] that states that if the surrogate loss function $\ell(\cdot,\cdot)$ being used is a Bregman divergence generated by a function that is C_S-strongly convex with respect to some norm

$\|\cdot\|$ then we can bound $\ell_{\text{NDCG}}(\mathbf{s}, \mathbf{r}) \leq \frac{C_F}{\sqrt{C_S}} \cdot \sqrt{\ell(\mathbf{s}, \mathbf{r})}$ where $C_F = 2 \left\| \left(\frac{1}{F(1)}, \ldots, \frac{1}{F(m)} \right)^\top \right\|_*$, F is the decay function used in the definition of NDCG and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. Note that we are using the "noiseless" version of the result where $r(\mathbf{x})$ is a deterministic function of $\mathbf{x}$.

In our case the squared loss is 2-strongly convex with respect to the L_2 norm which is its own dual. Hence $C_S = 2$ and $C_F = \mathcal{O}\left(\sqrt{\frac{m}{\log m}}\right)$, if $\hat{f} : \mathbf{x} \mapsto \langle \hat{\mathbf{w}}, \Psi(\mathbf{x}) \rangle$ is our final output, we get, for some constant C,

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\ell_{\text{NDCG}} \left(\hat{f}(\mathbf{x}), r(\mathbf{x}) \right) \right]\!\!\right] \leq C \sqrt{\frac{m}{\log m}} \cdot \sqrt{\epsilon_0 + 4Bm\epsilon_1} \leq C \sqrt{\frac{m}{\log m}} \cdot \sqrt{\epsilon_0} + C \frac{2m}{\sqrt{\log m}} \cdot \sqrt{B\epsilon_1}$$

which proves the claim. This affects the bounds given by Lemmata 15 and 16 since the dependence of the excess error on d and n will now be in terms of the inverse of their fourth roots instead of inverse of the square roots as was the case in regression and ordinal regression. □

We note that the (rather heavy) dependence of the final utility guarantee (that is $\mathcal{O}\left(\sqrt{m\epsilon_0}\right)$) on m is because the decay function $F(t) = \log(1 + t)$ chosen here (which seems to be a standard in literature but with little theoretical justification) is a very slowly growing function (it might sound a bit incongruous to have an increasing function as our *decay* function - however since this function appears in the denominator in the definition of NDCG, it effectively induces a decay). Using decay functions that grow super-linearly (or rather those that induce super-linear decays), we can ensure $\mathcal{O}\left(\sqrt{\epsilon_0}\right)$-usefulness since in those cases, $C_F = \mathcal{O}(1)$.

We next prove admissibility bounds for the ranking problem. The learning setting as well as the proof is different for ranking (due to presence of multiple entities in a single ranking instance), hence we shall provide all the arguments for completeness.

Theorem 31. *Every PSD kernel that is (ϵ_0, γ)-good for a ranking problem is also $\left(\epsilon_0 + \epsilon_1, \mathcal{O}\left(\frac{m\sqrt{m}}{\epsilon_1 \sqrt{\epsilon_1} \gamma^3}\right)\right)$-good as a similarity function for any $\epsilon_1 > 0$.*

Proof. For notational convenience, we shall assume that the RKHS $\mathcal{H}_K$ is finite dimensional so that we can talks in terms of finite dimensional matrices and vectors. As before, let $f(\mathbf{z}; \mathbf{W}) = \langle \mathbf{W}, \Phi_K(\mathbf{z}) \rangle$ and let $\mathbf{W}'$ be the minimizer of the following program.

$$\begin{aligned}
& \min_{\mathbf{W} \in \mathcal{H}_K} && \frac{1}{2} \|\mathbf{W}\|^2_{\mathcal{H}_K} + C \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\ell_{\text{sq}} \left(f(\mathbf{x}; \mathbf{W}), \eta(r(\mathbf{x})) \right) \right]\!\!\right] \\
\equiv & \min_{\mathbf{W} \in \mathcal{H}_K} && \frac{1}{2} \|\mathbf{W}\|^2_{\mathcal{H}_K} + C \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\sum_{i=1}^{m} \ell_{\text{sq}} \left(f(\mathbf{z}_i; \mathbf{W}), \eta(r(\mathbf{x}))_i \right) \right]\!\!\right] \\
\equiv & \min_{\mathbf{W} \in \mathcal{H}_K} && \frac{1}{2} \|\mathbf{W}\|^2_{\mathcal{H}_K} + C \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[\!\!\left[\ell_{\text{sq}} \left(f(\mathbf{z}_i; \mathbf{W}), \eta(r(\mathbf{x}))_i \right) \right]\!\!\right] \\
\equiv & \min_{\mathbf{W} \in \mathcal{H}_K} && \frac{1}{2} \|\mathbf{W}\|^2_{\mathcal{H}_K} + mC \mathop{\mathbb{E}}_{\mathbf{z} \sim \mu} \left[\!\!\left[\ell_{\text{sq}} \left(f(\mathbf{z}; \mathbf{W}), \tilde{r}(\mathbf{z}) \right) \right]\!\!\right] + C_{\mathcal{D}}
\end{aligned}$$

where for any $\mathbf{z} \in \mathcal{Q} \times \mathcal{P}$, $\tilde{r}(\mathbf{z})$ gives us the expected normalized relevance of this document-query pair across ranking instances and $C_{\mathcal{D}}$ is some constant independent of $\mathbf{W}$ and dependent solely on the underlying distributions. Using the goodness of the kernel K and the argument given in the proof of Lemma 17, it is possible to show that the vector $\mathbf{W}'$ has squared loss at most $\frac{1}{2C\gamma^2} + \epsilon_0$. Hence the only task remaining is to show that their exists a bounded weight function w such that for all $\mathbf{z} \in \mathcal{P} \times \mathcal{Q}$, we have $f(\mathbf{z}; \mathbf{W}) = \langle \mathbf{W}', \Phi_K(\mathbf{z}) \rangle = \mathop{\mathbb{E}}_{\mathbf{z}' \sim \mu} \left[\!\!\left[w(\mathbf{z}) K(\mathbf{z}, \mathbf{z}') \right]\!\!\right]$ which will prove the claim.

To do so we assume that the (finite) set of document-query pairs is $(\mathbf{z}_1, \ldots, \mathbf{z}_k)$ with $\mathbf{z}_i$ having probability μ_i and relevance $r_i = \tilde{r}(\mathbf{z}_i)$. Then the above program can equivalently be written as

$$\begin{aligned}
& \min_{\mathbf{W} \in \mathcal{H}_K} && \frac{1}{2} \|\mathbf{W}\|^2_{\mathcal{H}_K} + mC \sum_{i=1}^{k} \mu_i \ell_{\text{sq}} \left(\langle \mathbf{W}, \Phi_K(\mathbf{z}_i) \rangle, r_i \right) \\
\equiv & \min_{\mathbf{W} \in \mathcal{H}_K} && \frac{1}{2} \|\mathbf{W}\|^2_{\mathcal{H}_K} + mC \left\| \sqrt{P} X^\top \mathbf{W} - \sqrt{P} \mathbf{r} \right\|_2^2 \\
\equiv & \min_{\mathbf{W} \in \mathcal{H}_K} && \frac{1}{2} \|\mathbf{W}\|^2_{\mathcal{H}_K} + mC \left\| \tilde{X}^\top \mathbf{W} - \tilde{\mathbf{r}} \right\|_2^2 \\
\equiv & \min_{\alpha \in \mathbb{R}^{mn}} && \frac{1}{2} \|X\alpha\|^2_{\mathcal{H}_K} + mC \left\| \tilde{X}^\top X \alpha - \tilde{\mathbf{r}} \right\|_2^2
\end{aligned}$$

where $X = (\Phi_K(\mathbf{z}_1), \ldots, \Phi_K(\mathbf{z}_k))$, $\mathbf{r} = (r_1, \ldots, r_k)^\top$, P is the $k \times k$ diagonal matrix with $P_{ii} = \mu_i$, $\tilde{X} = X\sqrt{P}$ and $\tilde{\mathbf{r}} = \sqrt{P}\mathbf{r}$. The last step follows by the Representer Theorem which tells us that at the optima, $\mathbf{W}' = X\alpha$ for some $\alpha \in \mathbb{R}^k$.

Some simple linear algebra shows us that the minimizer α has the form

$$\begin{aligned}
\alpha &= \left(X^\top \tilde{X} \tilde{X}^\top X + \frac{1}{2mC} X^\top X \right)^{-1} X^\top \tilde{X} \tilde{\mathbf{r}} \\
&= \left(GPG + \frac{G}{2mC} \right)^{-1} GP\mathbf{r} \\
&= \left(PG + \frac{I}{2mC} \right)^{-1} G^{-1} GP\mathbf{r} \\
&= \left(PG + \frac{I}{2mC} \right)^{-1} P\mathbf{r}
\end{aligned}$$

where $G = X^\top X$ is the Gram matrix given by the kernel K. In the third step we have assumed that G does not have vanishing eigenvalues which can always be ensured by adding a small positive constant to the diagonal. Thus we have

$$\left(PG + \frac{I}{2mC} \right) \alpha = P\mathbf{r}$$

looking at the i^{th} element of both sides we have

$$\mu_i \sum_{j=1}^{k} \alpha_j K(\mathbf{z}_i, \mathbf{z}_j) + \frac{\alpha_i}{m2C} = \mu_i r_i$$

which gives us $\alpha_i = 2mC\mu_i \left(r_i - \langle \mathbf{W}', \Phi_K(\mathbf{z}_i) \rangle \right)$. Now assume, without loss of generality, that the relevance scores are normalized, i.e. $r_i \leq 1$ for all i. Thus we have

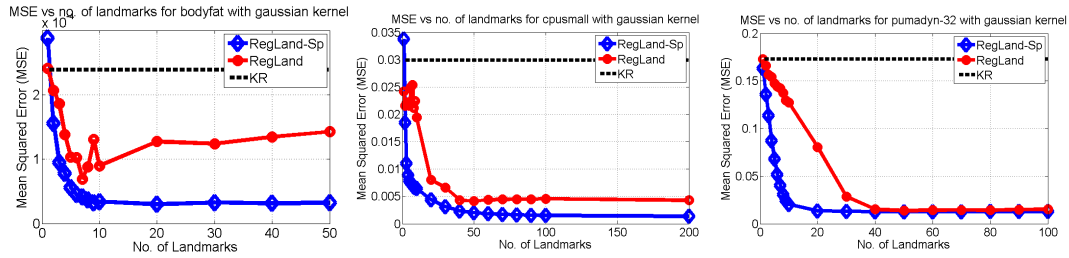
$$\frac{1}{2} \|\mathbf{W}'\|^2_{\mathcal{H}_K} + mC \left\| \tilde{X}^\top \mathbf{W}' - \tilde{\mathbf{r}} \right\|_2^2 \leq \frac{1}{2} \|\mathbf{0}\|^2_{\mathcal{H}_K} + mC \left\| \tilde{X}^\top \mathbf{0} - \tilde{\mathbf{r}} \right\|_2^2$$

which gives us $\frac{1}{2} \|\mathbf{W}'\|^2_{\mathcal{H}_K} \leq mC \|\tilde{\mathbf{r}}\|_2^2 \leq mC \sum_{i=1}^{k} \mu_i = mC$ which gives us $\|\mathbf{W}'\| \leq \sqrt{2mC}$. Since the kernel is already a normalized kernel, $\|\Phi_K(\mathbf{z}_i)\| \leq 1$ which gives us, by an application of Cauchy-Schwartz, $|\alpha_i| \leq 2mC\mu_i(1 + \sqrt{m2C}) \leq 5\mu_i mC\sqrt{mC}$.

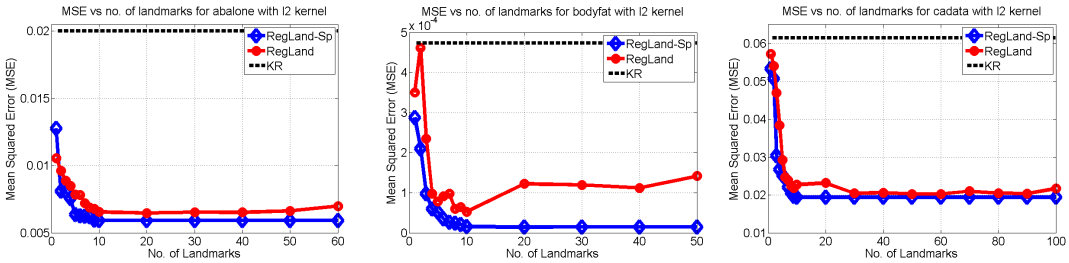
If we now establish a weight function over the domain $w_i = \frac{\alpha_i}{\mu_i}$, then $|w_i| \leq 5mC\sqrt{mC}$ and we can show that for all $\mathbf{z}$, we have $\langle \mathbf{W}', \Phi_K(\mathbf{z}) \rangle = \mathop{\mathbb{E}}_{\mathbf{z}' \sim \mu} [\![w(\mathbf{z}) K(\mathbf{z}, \mathbf{z}')]\!]$. Setting $C = \frac{1}{2\epsilon_1 \gamma^2}$ finishes the proof. □

Appendix G Supplementary Experimental Results

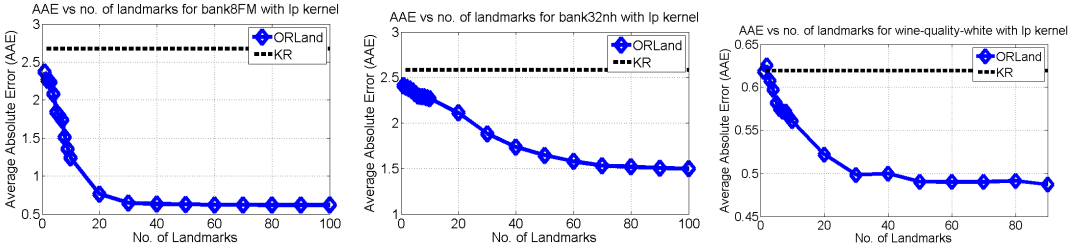
Below we present additional experimental results for regression and ordinal regression problems.



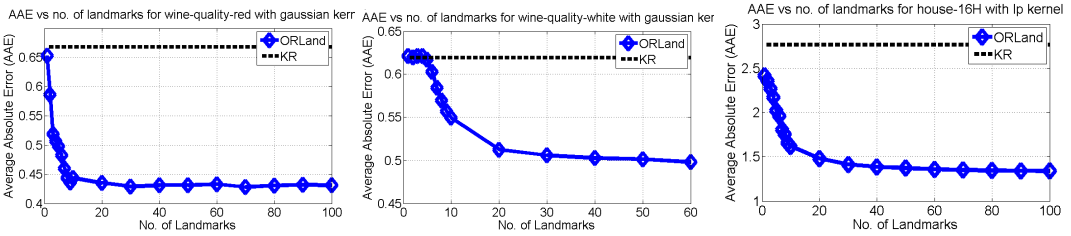
(a) Mean squared error for landmarking (RegLand), sparse landmarking (RegLand-Sp) and kernel regression (KR) for the Gaussian kernel



(b) Mean squared error for landmarking (RegLand), sparse landmarking (RegLand-Sp) and kernel regression (KR) for the Euclidean kernel



(c) Avg. absolute error for landmarking (ORLand) and kernel regression (KR) on ordinal regression datasets for the Manhattan kernel



(d) Avg. absolute error for landmarking (ORLand) and kernel regression (KR) on ordinal regression datasets for the Gaussian kernel

Figure 2: Performance of landmarking algorithms with increasing number of landmarks on real regression (Figures 2a and 2b) and ordinal regression datasets (Figures 2c and 2d) for various kernels.

G.1 Regression Experiments

We present results on various benchmark datasets considered in Section 4 for Gaussian $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}\right)$ and Euclidean: $K(\mathbf{x}, \mathbf{y}) = -\|\mathbf{x}-\mathbf{y}\|_2^2$ kernels. Following standard practice, we fixed σ to be the average pairwise distance between data points in the training set.

G.2 Ordinal Regression Experiments

We present results on various benchmark datasets considered in Section 4 for Gaussian $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}\right)$ and Manhattan: $K(\mathbf{x}, \mathbf{y}) = -\|\mathbf{x}-\mathbf{y}\|_1$ kernels.